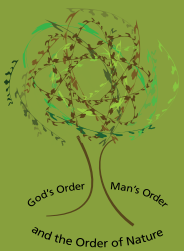
Evidence: For Policy

And Wheresoever Rigor is a Must

By Nancy Cartwright

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE



ISSN: 2045-5577
Order Project Discussion Paper Series

CONTENTS

ABOUT THIS VOLUME

This collection has been assembled under the auspices of the Templeton project, God's Order, Man's Order and the Order of Nature, which I helped direct. Much of my work on that project was concerned with Man's Order. Specifically, with how better to use scientific evidence to improve Man's Order. My research in this area grew out of a study of evidence-based policy, which attempts to use evidence from the sciences to evaluate whether policies that have been tried have succeeded and to predict whether those we are thinking of trying will produce the outcomes we aim for.

Since randomised controlled trials (RCTs) are widely deemed the gold standard for evidence in evidence-based policy, much of my Order Project work centred on what can and cannot be immediately inferred from positive results in a well-done RCT. You will see a discussion of the formalities in a number of the papers collected here. The point is to be clear what RCTs can rigorously show in order to get clear just what their results can be evidence for and what not – or better, what more needs to be the case to turn an RCT *result* into *evidence*. There are two further conditions, I argue, that are sufficient and more or less necessary if the same result (same "effect size") is to hold in a new setting as in the RCT study population, where "more or less necessary" just means that the same results could hold without these conditions but that would be pure serendipity. The two conditions are that the policy must play the same casual role in the target setting as in the study setting and that the study and target settings must have the same distribution (more properly, the same average) for the helping factors necessary for the policy to operate. Both these are explained in Section IV.

In working on the question "What makes an RCT result evidence for an effectiveness prediction?" in the context of the demands for rigour in evidence-based policy, it seems I have evolved an answer that works not just for RCT results and effectiveness predictions but across a wide range of result/hypothesis pairs where-ever a high premium is put on rigour. The answer picks up an old philosophical theme, that evidence is not a 2- but rather a 3-place relation. E is evidence for hypothesis H *relative* to something else. What else? I propose a demanding answer: a sound argument for H that uses E as an essential premise. This work on a theory of evidence where rigour is required is the most recent I have done in this project and so is still in development. I discuss this account of evidence in Section III.3 and again in parts of IV.1. My thinking here has been much helped by grappling with the exciting work on evidence by Sherilyn Roush, trying to understand which of her ideas can carry over to issues like those in evidence-based policy and which not. You can read about that in a joint paper I wrote with Damien Fennell, which is reproduced here as Section III.1.

The papers in Section I introduce my project. Those in Section II and III then split in two directions. The first route, which constitutes Section II, begins concretely with the current context in which RCTs are held as the highest form of evidence for predictions about policy effectiveness. The papers in this section bring to light some deep problems with this standard view. The second route, examined in Section III, begins more abstractly and considers what an adequate theory of evidence could consist in for evidence-based policy and anywhere else where rigour matters. Finally the two routes rejoin in Section IV, on the role of evidence in policy.

The collection here presents a sample of the interlocking work on evidence that I have done during the Order Project. You can find a list of further work below. Some of the basic points on which the arguments build necessarily appear in more than one paper. I apologise to readers of this volume for the repetition.

I am very grateful to the Templeton Foundation for making this work possible and for supporting the exciting research team with whom I have worked on it. I think everyone involved has learned and benefitted from our interactions and interchanges on all aspects of the Order Project. I would also like to thank the Spencer Foundation, the British Academy, LSE's Grantham Research Institute on Climate Change and the UCSD Faculty Senate for support for aspects of the research reported here.

Nancy Cartwright

FOR MORE DETAILS

The papers collected here are a sample from Cartwright's recent work on evidence and evidence-based policy. The basic points presented in this volume are developed and explored in more detail in a number of further works.

Further aspects of the discussion of RCTs from Section II are examined in

Cartwright, Nancy 'What is This Thing Called Efficacy' in *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*, Mantzavinos, C. (ed.), Cambridge: Cambridge University Press, 2009, pp. 185-206.

Cartwright, Nancy 'What Are Randomized Controlled Trials Good For?' *Philosophical Studies 147* (2010): 59-70.

Cartwright, Nancy and Munro, Eileen 'The Limitations of Randomized Controlled Trials in Predicting Effectiveness', *Journal of Evaluation in Clinical Practice* 16.2 (2010):260-266.

Cartwright, Nancy 'Predicting "It Will Work for Us": (Way) Beyond Statistics', in *Causality in the Sciences*, Illari, P.M., Russo, F., and Williamson, J. (eds.), New York: Oxford University Press, 2010, 750-768.

Cartwright, Nancy 'Knowing What We Are Talking About: Why Evidence Doesn't Always Travel' *Evidence and Policy* 9.1 (2013): 97-112.

Additional aspects of Section III are explored in

Cartwright, Nancy 'Evidence-Based Policy: What's To Be Done About Relevance', in *Philosophical Models, Methods, and Evidence: Topics in the Philosophy of Science.*

Proceedings of the Thirty-Eighth Oberlin Colloquium in Philosophy. Published in *Philosophical Studies* 143 (2009): 127-136.

Cartwright, Nancy 'Predicting What Will Happen When We Act: What Counts as Warrant?' *Preventive Medicine* (53), September 2011, 221-224.

Further aspects of Section IV are explored in

Cartwright, Nancy and Stegenga, Jacob. 'A Theory of Evidence for Evidence Based Policy' in *Evidence, Inference and Enquiry*, Dawid P. Twining, W. and Vasilaki, M. (eds.), New York: Oxford University Press, 2012.

LOCATIONS OF ORIGINAL PUBLICATION

Chapter 1.1 'Evidence-Based Policy: Where is Our Theory of Evidence?' was published in the *Journal of Children's Services* (special edition), December 2009 4(4): 6-14.

Chapter 1.2 'Evidence for Evidence Based Policy' was presented at the Home Office Seminar on Criminology and Evidence-Based Policy June 10, 2008.

Chapter 2.1 'A Philosopher's View of the Long Road from RCTs to Effectiveness' was published in *The Lancet* (Art of Medicine Section), 377 (2011): 1400-1401.

Chapter 2.2 'Are RCTs the Gold Standard?' was published in *BioSocieties* 2 (2007): 11-20.

Chapter 2.3 'RCTs, Evidence and Predicting Policy Effectiveness' was published in The Oxford Handbook of the Philosophy of the Social Sciences, Kincaid, H. (ed.), New York: Oxford University Press, 2012, pp. 298-318.

Chapter 3.1 'Does Roush show that Evidence Should be Probable?' (with Damien Fennell) was published in *Synthese* 175 (2010): 289-310.

Chapter 3.2 'Evidence, External Validity and Explanatory Relevance' was published in Philosophy of Science Matters: The Philosophy of Peter Achinstein, Morgan, G. (ed.), New York: Oxford University Press, 2011, 15-28.

Chapter 3.3 'Evidence, Argument and Prediction' was presented at the European Philosophy of Science Association 2011 and will be published in *Proceedings of the European Philosophy of Science Association (Forthcoming in V. Karakostas and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings 2)*

Chapter 4.1 'The Theory that Backs up What We Say' (with Jeremy Hardie) was published as chapter I.B in Evidence Based Policy: A Practical Guide to Doing it Better, New York: Oxford University Press, 2012, pp. 14-58.

Chapter 4.2 'Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps' was the Presidential Address of the Philosophy of Science Association, 2010. Published in *Philosophy of Science* 79.5 (2012): 973-989.

SECTION I

EVIDENCE-BASED POLICY: SO, WHAT'S EVIDENCE?

1.1 EVIDENCE-BASED POLICY: WHERE IS OUR THEORY OF EVIDENCE?

(From *The Journal of Children's Services,* 2009*)*

Nancy Cartwright with Andrew Goldfinch and Jeremy Howick.

The rise of evidence-based policy

In both the UK and the US there is an increasing drive to use evidence to inform, develop and refine policy and practice. This push to improve how research and analysis informs policy and practice is increasingly being felt in a wide range of areas: in addition to evidence-based health and social care, we now hear of evidence-based housing policy, transport policy, education and criminal justice. Since the election of the Labour Government in 1997, the UK has been firmly committed to evidence-based policy as a way of developing social programmes. The UK Government signalled its commitment to evidence-based policy in the 1999 White Paper *Modernising Government*, which calls for the "better use of evidence and research in policy-making and better focus on policies that will deliver long term goals" and stipulates evidence as a key principle for policy making (Cabinet Office, 1999: 16). A year later, the Cabinet Office's Performance and Innovation Unit (2000) called for a "fundamental change in culture" in order to place good analysis at the centre of policy-making and recommended that training for new Ministers and senior civil servants "should emphasise the importance of analysis for evidence-based policy" (p 4). In response to this recommendation the UK's National School of Government, which provides training for the civil service, now runs regular courses on analytical skills and evaluation methods, including introductions to, and overviews of, evidence-based policy making.

An example of evidence-based approach to policy making is the UK Sure Start programme. Initiated in 2001, the aim of the programme is to break the cycle of poverty by providing children and families with childcare, health and educational support. The Sure Start programme has been evidence-based from the start, using extensive reviews of research findings on what approaches and early interventions are most likely to work; its execution and continuing evaluation and refinement have also been evidence-based (Hunter, 2003). Another notable example is the UK's National Institute for Clinical Excellence (NICE),[1] which provides regulatory guidelines for the National Health Service (NHS) on particular treatments. These guidelines are based on reviews on the effectiveness and cost-effectiveness of various treatments.

In the US, the Department of Education is actively committed to furthering evidence-based approaches to education policy and practice. The Department's Institute of Education Sciences established the What Works Clearinghouse in 2002 "to provide educators, policymakers, researchers, and the public with a central and trusted source

[1] www.nice.org.uk/

of scientific evidence of what works in education"[2]. Furthermore, the Department in 2005 implemented a recommendation by the Coalition for Evidence-Based Policy[3] that projects that include a randomised evaluation should have priority in its grant process.

The commitment to evidence-based policy has been matched with funding. In June 2000, the UK Treasury established the Evidence-Based Policy Fund. With a budget of £4 million over two years, the aim of the fund was to support cross-cutting research and links between research institutes, universities, and government. Several government departments have also contributed funding to the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre).[4] Based at the Institute for Education, University of London, the EPPI-Centre collects, reviews and organises the results of evidence-based public policy and research in an accessible way for policy-makers and others. The Economic and Social Research Council (ESRC) funded the UK Centre for Evidence Based Policy, which was based at King's College London. The centre co-ordinated a network of research centres dedicated to promoting evidence-based policy and practice by contributing to the development of methods for evaluating and summarising research.

Not only are evidence-based approaches to policy making funded by governments but also some government funding is increasingly being tied to demands for evidence. For example, proposals to expand the Sure Start programme led to a £16 million research project to establish whether the programme was achieving results (see Belsky *et al*, 2007; Melhuish *et al*, 2008). In the US, the so-called No Child Left Behind Act 2001 enshrines in law the principle that federal funds should support educational activities that are based on "scientifically-based research". Title I funding is designed to help schools improve the achievement of disadvantaged students. Those schools that receive Title I funding are required by the Act to use effective methods and strategies grounded in scientifically-based research.

In addition to executives, legislatures too are beginning to take a strong interest in evidence-based approaches to policy making. In November 2005, the UK Parliament's Select Committee on Science and Technology agreed to establish an inquiry on "Scientific Advice, Risk and Evidence: How Government Handles Them". The inquiry examined the extent to which polices are evidence-based, what mechanisms are in place for the use of evidence, and the way in which guidelines relating to the use of advice are being applied. Issues addressed include "sources and handling of advice" and "the relationship between scientific advice and policy development". Particular questions explored by the Select Committee include: "What mechanisms are in place to ensure that policies are based on available evidence?"; "Are departments engaging effectively in horizon scanning activities and how are these influencing policy?"; and "Is Government managing scientific advice on cross-departmental issues effectively?"

Although the drive for evidence-based policy is strongest in the UK and US, the movement is picking up elsewhere in Europe. In its 2001 white paper on governance, the European Union acknowledged that,

"Scientific and other experts play an increasingly significant role in preparing and monitoring decisions. From human and animal health to social legislation, the Institutions rely on specialist expertise to anticipate and identify the nature of the problems and uncertainties that the Union faces, to take decisions and to ensure that risks can be explained clearly and simply to the public." (Commission of the European Communities, 2001)

So in the UK, the US, and gradually in Europe, at the executive and legislative levels, and pushed by national and international organisations such as the Campbell and Cochrane Collaborations,[5] institutions and regulations are increasingly attempting to ensure that evidence is appropriately considered at various levels of decision-making processes.

Evidence: the missing theory

Evidence-based policy is on the rise then, and all to the good we should suppose. *Except* that we do not have a theory of evidence that can be called upon in policy deliberations. We are supposed to base our policies on evidence but how exactly are we to proceed: what is to count as evidence and how shall we use it? My central thesis is that we lack a practicable theory of evidence – one that can be put to use for evidence-based policy. There are three essential ingredients missing. We do not have:

A reasonable and practicable concept of evidence

A reasonable and practicable account of what different pieces of evidence say about a hypothesis and with what strength they speak (see Hammersley, 2005 for a discussion of the variety of kinds of questions evidence can speak to)

A reasonable and practicable account of how to evaluate a hypothesis in the light of all the candidate evidence.

Philosophical accounts of the concept of evidence

What is it in virtue of which a fact is evidence for a hypothesis? Our philosophical accounts fall into two categories. First are accounts based on some features of the probabilistic relations between the evidence and the hypothesis – for example, increase in probability or various functions of likelihoods (see Mayo, 1996 Chapter 3 for an overview of such positions). These are not useful for evidence-based policy. What we need is a concept of evidence that we can use to judge whether some fact should be taken into consideration – whether it should be "on

[2] www.whatworks.ed.gov/whoweare/overview.html

[3] http://coalition4evidence.org/wordpress/

[4] http://eppi.ioe.ac.uk/cms/

[5] www.cochrane.co.uk and www.capmbellcollaboration.org.

the table" for consideration. Then we would expect to look at all the evidence on the table to decide on the probability of the proposed policy claim. Concepts of evidence based on facts about probabilities put the cart before the horse. We need a concept that can give guidance about what is relevant to consider in deciding on the probability of the hypothesis not one that requires that we already know significant facts about the probability of the hypothesis on various pieces of evidence.

Second are those accounts that are based on facts about explanation – for example, versions of inference to the best explanation (Lipton, 2004) or explanatory connectedness (Achinstein, 2001). The problem here is the concept of explanation. A good many accounts end up explaining explanation by reference to probability relations between the "explanans" [the means of making plain] and the "explanandum" [that which is being made plain]. This simply recreates the previous problem. Also, it seems to me that the concept is too narrow. Suppose for example that we are considering a policy to combat segregation, perhaps making "diversity training" mandatory in schools. But recall Thomas Schelling's (1978) game-theory model where checkers are moved on a checkerboard so as to avoid any one checker being the only one of its colour in a group. Eventually clumping occurs even though no moves are designed to put checkers in neighbourhoods that are predominately of their own colour. This is an important model to consider in judging the efficacy of the program for diversity training in reducing segregation.

But it is far-fetched to see it as explanatorily connected with the claim that the policy will be efficacious.

Besides these problems, our accounts of evidence also tend to be accounts of *genuine evidence*. But we need an account of what makes something *candidate evidence*. I think I can convince you that you have such a concept by pointing out that we are often ready to blame people for failing to report facts that, though they may turn out *not* to be evidence, under some scenarios *could have been*. Mystery stories are rife with examples. In these cases the aim is to evaluate a retrospective rather than a prospective causal claim but the point is the same.

Hypothesis: John Jones killed Roger Ackroyd. He could have done so by doing A, B and then C. Ah, but he couldn't because in that case he would have had to travel between Binsey and Summertown in 8 minutes and even the fastest car could not do that. But you are Jones's girlfriend and you know he keeps a fast cross-country motorcycle in his garage so he could have gotten there across Port Meadow in time. You are blameworthy if you do not speak up. Yet if it turns out that A, B and C did not occur, the fact that he owns a dirt bike is totally irrelevant to the hypothesis that Jones caused Ackroyd's death. The case would be exactly similar if you were on a commission and did not report some fact you knew that might be relevant to the efficacy of a policy under consideration but in the end turns out not to be.

What we are urged to do in practice

We also have philosophical accounts that provide the second and third components that I claim to be missing from our theory of evidence. The problem with these is the same as with our philosophical accounts of what evidence is: they are generally not very practicable. They are well reasoned and make sense. But they are usually either too abstract or too circular to provide useable advice about how to conduct evidence-based policy. By contrast, there are now available a host of far more usable schemes – *evidence-ranking schemes*. The problem is that these schemes are not well reasoned and sensible; many seem to me to be daft, indeed pernicious. Yet they are being pushed by a number of influential institutions, not the least of which are the UK and US governments.

These schemes provide all three of my "missing" components in one fell swoop. Kinds of evidence are ranked according to their "quality". Then: (1) Evidence is all and only facts of the kind listed in the ranking. (2) All evidence is taken to speak for or against the *truth of an hypothesis* and the strength of its support is in line with its quality: top ranked evidence indicates that the hypothesis is very likely true, and as quality decreases, so does the strength of support for the truth of the hypothesis. (3) In general the recommendations associated with these schemes do not combine evidence at all.

Very often the advice is: if you have top grade evidence, go with what that says. The US Department of Education, for instance, which requires evidence of efficacy in order for a school to receive Title 1 support, tells us that RCTs (randomised clinical trials) are needed to establish strong evidence and that "Two or more typical school settings, including a setting similar to that of your schools/classrooms" is the quantity of evidence needed.

There are a vast number of similar schemes available. I choose as an example one particularly thoughtful one, SIGN (Scottish Intercollegiate Guideline Network).[6] As their own document reports:

"SIGN formerly used the levels of evidence developed by the US Agency for Health Care Policy and Research (AHCPR, now the US Agency for Health Research and Quality, AHRQ). However as a number of limitations were becoming apparent in that system, a review was carried out and new levels of evidence and associated grades of recommendation were developed. Following extensive consultation and international peer review, the new grading system was introduced in Autumn 2000." (SIGN, 2008: 36)

[6] www.sign.ac.uk

The SIGN grading system is this:

SIGN (Scottish Intercollegiate Guideline Network) grading system

Levels of evidence:

1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1 -	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2++	High quality systematic reviews of case-control or cohort studies High quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal
2 -	Case control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal
3	Non-analytic studies, eg. case reports, case series
4	Expert opinion

The grading scheme goes like this:

Grades of recommendation:

A	At least one meta analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
B	A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 1++ or 1+
C	A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 2++
D	Evidence level 3 or 4; or extrapolated evidence from studies rated as 2+

Look now at samples of the kind of advice on offer about how to arrive at an overall judgment:

Statements that one piece of "level 1++" evidence is sufficient:

GRADE Working Group: "Once the results of high quality randomized trials are available, few people would argue for continuing to base recommendations on non-randomised studies with discrepant results" (Atkins *et al*, 2004: 2). [GRADE (Grades of Recommendation Assessment, Development and Evaluation) is an international project aimed at developing a methodologically sound system that can be applied across countries and cultures[7].]

SIGN: The following quote from the SIGN 50 document seems to imply that if there are RCTs, the other evidence need not be considered: "It is also intended to allow more weight to be given to recommendations supported by good quality observational studies where RCTs are not available for practical or ethical reasons." (SIGN, 2008: 36)

EBM [Evidence-based medicine]: "If the study wasn't randomized, we'd suggest that you stop reading it and go on to the next article in your search" (Sackett *et al*, 2000: 108).

Cochrane Collaboration: In personal correspondence with Jeremy Howick [co-author], Julian Higgins of the Cochrane Collaboration replied to the question of whether evidence from RCTs is sufficient, with the following statement: "I'm sure there are very many people who subscribe to this view [that RCT evidence is sufficient] (if interpreted as further evidence on the same questions that the RCTs address). Indeed, one might infer this from the fact that the majority of Cochrane reviews include only RCTs. This strongly implies that the authors believe there is no need to look at other evidence (or believe that "Cochrane" thinks they shouldn't look at other types of evidence). I have much sympathy with this, given the numerous unpredictable and largely poorly understood biases in observational studies."

In answer to the question of whether a single well-done RCT trumps evidence from any number of observational studies, Julian Higgins states that "If the RCT was done well, then I would always claim this is either the right answer or the answer to a different question from the observational studies." [The Cochrane Collaboration is dedicated to encouraging RCTs]

So, what's wrong with that?
Virtually everything.

1. The concept of evidence involved is too restrictive

Hardly anything gets on the table. This is bad for a number of different reasons. To start with, the type of evidence restricts the type of conclusion for which we can have evidence. These schemes are all for judging

[7] www.gradeworkinggroup.org

efficacy claims. But more, concepts in the study have to match exactly with those in the policy claim; especially they must be completely operationalisable and they must be operationalised in the same way. How could Oxford Council have used evidence like this to decide whether to build a leisure centre in the new housing estate at Blackbird Leys? Certainly not in the way envisaged in the grading schemes. Candidate evidence is not even in the ballpark.

The advantage of an RCT is that it can *clinch* results. If the ideal conditions for an RCT are met, positive results *deductively imply* that in at least some subpopulations of the experimental population the treatment causes the relevant effect. But other methods have this advantage as well and they are not in the list. These include various econometric modelling techniques, deduction from established theory and experiments in ideal model systems (Cartwright, 2007a: Section I.3).

A host of other methods that can *vouch for* results even if not clinch them are excluded. These include the hypothetico-deductive method when used for confirmation, qualitative comparative analysis, game-theory modelling, ethnographic methods, and so on. Moreover, any "voucher" can be turned into a "clincher" by adding some additional premises – premises that may be reasonable to entertain in particular cases. All methods presuppose other assumptions. These ranking schemes seem to presuppose that the background assumptions required by the methods listed are more likely to be true for all cases than those for methods omitted, which is highly implausible.

2. The claims about strength of evidence in the rankings are mistaken

Much is written about the pros and cons of the specific kinds of evidence that appear in these listings – fully randomised trials, partially randomised trials, observational studies, and so forth. I want to concentrate instead on the basic underlying ideas, which I think are way off base. I have already noted that the kinds of evidence permitted are only good for efficacy claims so I shall confine my attention to these, ignoring other policy issues such as claims about side effects (which nevertheless turn out to be an important issue in the example I will use), about implementation, about the effects of moral, cultural and political considerations, about estimates of costs and the like. I shall also concentrate on RCT evidence for concreteness but what I say can be carried over, mutatis mutandis, to other types of evidence that these rankings admit.

Consider: we wish to evaluate a proposal to do A in order to achieve R: say to treat African children who are HIV-infected prophylatically with an inexpensive antibiotic called "cotrimoxazole" in order to reduce mortality and morbidity from opportunistic disease until they are old enough for retroviral treatment, as in the 2005 UNAIDS and UNICEF call to ensure that prophylaxis with cotrimoxazole reaches 80 per cent of children in need by 2010 (UNAIDS, 2006: 165). An evidence ranking scheme tells us which kinds of evidence speak strongly for or against this proposal, which less strongly. In this case the justification for the policy is an RCT on children in Zambia published in the *Lancet* in 2004, which concluded that the antibiotic reduced mortality in HIV-infected children by more than 40 per cent (Chintu *et al*, 2004: 1870).

What is the underlying logic that shows how a study like this – assuming even that it meets all the ideal requirements – can serve as strong evidence for the efficacy of the policy? As far as I can see the most plausible construction of the underlying justification assumes that actions are justified by principles.

We suppose:
(a) There is a certain type of HIV-infected child population, *T*, for which the Zambian RCT establishes "In T cotrimoxazole reduces average morbidity/mortality".
(b) The target population – in this case HIV-infected children in resource-poor settings across Africa – is of type T.
(c) So administering cotrimoxazole in the target population will reduce average morbidity/mortality.

That is, we need some way to get from the evidence to the conclusion, and a way that shows how this evidence can speak so strongly for the conclusion. I think the only way it can work is via an intermediate principle. But this won't do since both the way up to the principle and the way back down to the policy are shaky, and for much the same reason: how to specify T. This is now explained in more detail.

As regards moving from principle to policy, what is wrong here is what is generally wrong in supposing you can read off conclusions about single cases from scientifically established principles: almost all principles are defeasible and those that are not (like "All men are mortal") do not provide very detailed advice. We can all imagine a vast variety of happenings that can defeat the policy efforts even in the face of the principle. One may have the happy idea that if the target population is really of the right kind – kind T, whatever that is – the defeaters will be distributed the same in the target as in the trial population so the conclusion will still obtain. That has its own problems:

We do not know what T is. This means that the guidelines may be able to provide sound advice but it is not practicable advice: we do not know how to tell whether we are following it or not. Our target population may start out satisfying the characterisation "T", whatever that is, but our efforts to implement the policy may change the distribution of defeating conditions or the underlying causal structure. This is a common worry about interventions in economics (Lucas, 1976, 1988) but not much discussed in the evidence-ranking and grading schemes.

In relation to moving from RCT to principle, a positive result in an ideal RCT can establish that in at least one subpopulation of the population involved in the trials the treatment causes the relevant effect. It can also establish that the average result in this population is improvement in the effect. The principle says the treatment causes the relevant effect, or produces an average improvement, in any population of type T. How do we get from

the first to the second? Laying aside Hume's problem of induction,[8] we suppose that the positive result will hold in any population like the one in the trial. Hence the emphasis on identifying T: "like" in what respects?

This is obviously not an unfamiliar problem. We do of course pay attention to what constitutes T. For instance, there were earlier RCTs in Cote d'Ivoire involving the treatment of adults with cotrimoxazole (Wiktor *et al*, 1999) These obviously were not good enough because a population of children can be very different from one of adults. Moreover, many African children live in areas with high rates of bacterial resistance. So the RCT that is used to justify the UNAIDS and UNICEF-proposed policy was performed on children and in Zambia where there are high rates of bacterial resistance to cotrimoxazole. But what else might be relevant?

The answer is a tough one. The only way to characterize T that works is – "populations that have just the same causal structure and the same joint probability distribution across all relevant variables as the population in the study". (This is clearly not really true. But this is the only characterisation that does not depend on details about what the probability distribution or causal structure are – see Cartwright, 2007b for a fuller discussion.) And this is clearly not a practicable description. We can try to sidestep the problem by insisting that the experimental population be a random sample from the target. How practicable is that, say for our cotrimoxazole policy? Moreover, random sampling procedures require a great deal of knowledge of the relevant structure of the population sampled if they are to be at all reliable. Not only do we not in general have such knowledge – the guidelines generally do not take this much into account.

My basic point here is much the same as the one I made in discussing clinchers and vouchers. Nothing can count as evidence for anything except relative to a host of auxiliary assumptions; and the strength with which a body of evidence supports a hypothesis can never be higher than the credibility of these auxiliaries. The privileged items that tend to appear in evidence-ranking schemes have built-in methods for assuring that a few of the necessary auxiliaries are met – blinding in RCTs, for example, is good at ensuring that one source of confounding for the results is eliminated. But there are huge gaps left. And there is no reasonable promise that the gaps are in general smaller than with types of evidence that are commonly not even allowed on the table by these schemes.

3. The advice about how to combine evidence is dreadful

Grading schemes do not combine evidence at all – they go with what is on top. But it seems to me to be daft to throw away evidence. Why are we urged to do it? Because we do not have a good theory of exactly why and how different types of evidence are evidence and we do not have a good account of how to make an assessment on the basis of a total body of evidence. Since we lack a prescription for how to do it properly, we are urged not to do it at all. That seems daft too. But I think it is the chief reason that operates. That is why the philosophical task is so important.

[8] The problem of induction is the philosophical question of whether inductive reasoning – ie. making a series of observations and inferring a new claim based on them – leads to knowledge.

Conclusion

We need to develop a practicable theory of evidence, a theory that will work for evidence-based policy. But it had better be a good theory, one that is both sound and usable: that is, a theory that is both practicable and philosophical.

References

Achinstein, P (2001) *The Book of Evidence*, Oxford: Oxford University Press.

Atkins D, Best D, Briss PA *et al* [GRADE Working Group] (2004) Grading quality of evidence and strength of recommendations, BMJ 328 (7454):1490 (19 June), doi:10.1136/bmj.328.7454.1490.

Belsky, J., Barnes, J. & Melhuish, E. (eds) (2007) *The National Evaluation of Sure Start: Does Area-based Early Intervention Work?*, Bristol: The Policy Press.

Cabinet Office (1999) *Modernising Government*, White Paper Cm 4310, London: HMSO.

Cabinet Office Performance and Innovation Unit (2000) *Adding It Up: Improving Analysis & Modelling in Central Government*, London: HMSO.

Cartwright, N. (2007a) *Hunting Causes and Using Them: Approaches in Philosophy and Economics*, Cambridge: Cambridge University Press.

———. (2007b) 'Are RCTs the gold standard?', *BioSocieties* 2, 11-20.

Chintu C, Bhat GJ, Walker AS, *et al* (2004) 'Co-trimoxazole as prophylaxis against opportunistic infections in HIV-infected Zambian children (CHAP): a double-blind randomized placebo-controlled trial.' *Lancet* 364, 1865-71.

Commission of the European Communities (2001) *European Governance: A White Paper*, COM(2001) 428 final, Brussels: Commission of the European Communities.

Hammersley M (2005) Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for

research-based policymaking and practice, *Evidence and Policy*, 1 (1) 85-100.

Hunter DJ (2003) 'Evidence-based policy and practice: riding for a fall?', *Journal of the Royal Society of Medicine* 96(4), 194-196.

Lipton, P (2004) *Inference to the Best Explanation*, London: Routledge.

Lucas, RE (1976) Econometric policy evaluation: a critique, in Lucas, RE (ed) (1981) *Studies in Business Cycle Theory*, Oxford: Basil Blackwell.

——— (1988) 'On the Mechanics of Economic Development', *Journal of Monetary Economics*, 22, 3-32.

Mayo, D (1996) *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Melhuish E, Belsky J, Leyland A *et al* (2008) 'A quasi-experimental study of effects of fully-established Sure Start Local Programmes on three-year-old children and their families', *The Lancet*, 372, 1641-1647.

Sackett DL, Straus SE, Richardson WS, Rosenberg & Haynes RB (2000) '*Evidence-Based Medicine: How to Practice and Teach EBM* (Second Edition), Edinburgh: Churchill Livingstone'.

Schelling TC (1978) *Micromotives and Macrobehaviour*, New York: Norton.

SIGN (Scottish Intercollegiate Guidelines Network) (2008) *SIGN 50: A Guideline Developer's Handbook (Revised edition, January 2008)*, Edinburgh; SIGN Executive.

UNAIDS (2006) *2006 Report on the Global AIDS Epidemic: Executive Summary*, / UNAIDS, available online at: http://data.unaids.org

Wiktor SZ, Sassan-Morokro MD, Grant AD *et al* (1999) 'Efficacy of trimethoprim-sulphamethoxazole prophylaxis to decrease morbidity and mortality in HIV-1-infected patients with tuberculosis in Abidjan, Côte d'Ivoire: a randomised controlled trial.' *Lancet*, May 1, 353 (9163): 1469-75.

1.2 EVIDENCE FOR EVIDENCED-BASED POLICY

(Presented originally at a UK Home Office Seminar on Criminology and Evidence-Based Policy, June 2008.)

I start by announcing that I am firmly in favour of evidence-based policy, despite the disappointments with it. Though evidence is not the only consideration, it is clearly better to look at the evidence and think hard about it than not. I am not at all surprised however that it has had disappointing results in both the UK and the US since we are not giving good advice about how to go about it.

Two lost questions

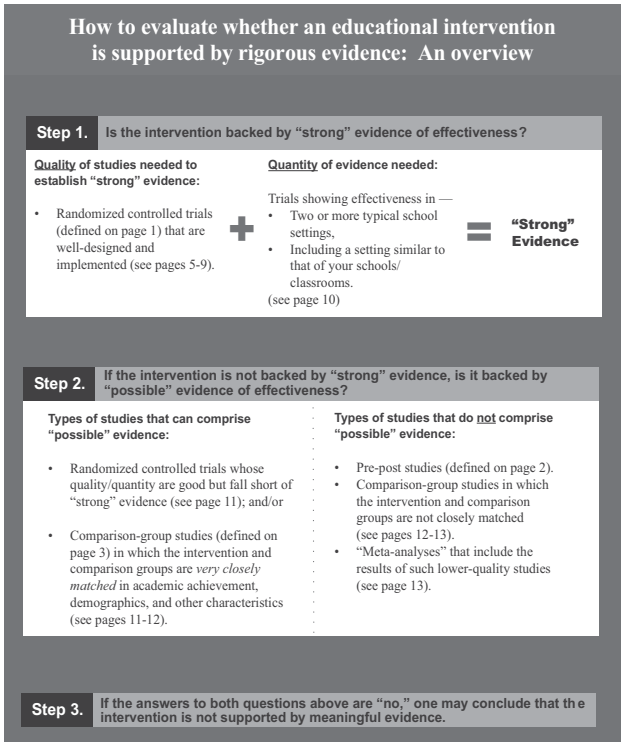
For today I stick with the issue of evidence for effectiveness; that is, for the claim that a policy will produce the desired ends when implemented – how, when and where it will be implemented. I start with a truism: In evaluating the effectiveness of a proposed policy we need credible evidence that speaks for or against the policy and we need to know what to do with the evidence when we have it. This truism naturally generates three questions:

1. What counts as *credible* evidence – what evidence claims are likely to be true?
2. What evidence claims are *relevant* – what claims speak for or against the proposed policy and how strongly?
3. How should the evidence be *integrated* – how should the probability of the policy being effective be established in light of all the evidence?

Begin with *credibility*. We do not wish to enter claims into the record of evidence that are themselves not likely to be true. Compare: In deciding if a person is guilty or innocent we do not take into account the testimony of a witness without good reason to think the witness is telling the truth. So too in deciding if a policy will be effective or not.

Suppose however just for the moment that this is no issue. Suppose we have at our disposal the entire encyclopaedia of unified science, an encyclopaedia that contains within it all the true claims there are. But for deliberating about a particular policy we are not going to cart the entire encyclopaedia to the table. Rather we want a selection – we want on the table only true facts that are *relevant* to the effectiveness of the policy. And given a collection of relevant facts we want to know how to assess the probability that the policy will be effective in light of them. How are we supposed to make all these decisions?

Let us look at an American case that is particularly egregious. It does not have to do with crime but with education: the so-called "No Child Left Behind" legislation. I use it because it illustrates my points clearly and sharply. Go to the US Dept. of Education website, as school masters are supposed to do, and this is the advice you find:



[From Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide, December 2003. US Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance.]

The left-hand side of the + sign – "quality" of evidence – plus all of step 2 is in answer to the first question: when is a particular kind of evidence claim credible? The US Dept of Education's is typical of a large number of evidence-ranking schemes currently available, including the Maryland rules. These schemes gauge the credibility of causal claims based on their associated research design. The advice they give is excellent, and detailed. As we know, there are very long checklists of demands a trial must meet to earn the label "well-designed and well-implemented".

Questions 2 and 3 are lumped together under the heading "quantity of evidence". The advice on relevance, Question 2, is short indeed – the RCT should be in "settings similar to that of your schools/classrooms" and page 10, which this page refers to, adds only 4 lines describing one case – trials on white suburban populations do not constitute strong evidence

for large inner city schools serving primarily minority students. This is like the warning: Beware, results about police-initiated fear reduction programmes in large estates in South Birmingham and Southwark may not carry over to Belgrave Square.

The answer to question 3 is equally short: two positive relevant RCTs are strong evidence for effectiveness. So with two good RCTs we can assign a high probability that the policy will work in our school.

One obvious problem is that this violates the principle of total evidence, which is at the heart of practice in the natural sciences: look at all the evidence, strong and weak, of various kinds and from various sources. Other guidelines do better in this regard. For instance the NICE guidelines allow us to consider all the evidence, weak and strong, pro and con. But we are still without advice about how to go about considering it.

The well-known Maryland rules prominent in criminology are very much the same. Here is the entire section headed "*integration* of evidence". This is where we should look for an answer to question 3 – how should we combine the evidence to evaluate the probability that the policy will be effective when we implement it? The integration offered is no integration at all. We are told to judge that a programme "works" if it has positive results in two rigorous studies, with an acknowledged need for "judgement" in the end:

Integration of Evidence. The end product of the analysis of empirical evidence contains a range of findings with respect to effectiveness. In the interests of clarity of presentation for policy analysis purposes, we organize the presentation of material in each chapter by the content of the findings, rather than the priority of the program. The content is defined both the strength of the scientific evidence and the strength (and direction) of the program effects. We ultimately report on four categories of effectiveness.

Program categories are sorted into these effectiveness categories using the following rule:

Works (1): At least two studies with methodological rigor greater than or equal to "3" reporting significance tests have found crime prevention effects for the program condition, and where effect sizes are available, the effect is at least one-tenth of one standard deviation (eg, effect size = .1) better than the effects for the control condition, and the preponderance of the evidence supports the same conclusion.

Doesn't Work (2): At least two studies with methodological rigor greater than or equal to "3" reporting significance tests have found no effect favoring the program condition, and the preponderance of other evidence supports the same conclusion.

Promising (3): At least one study with methodological rigor greater than or equal to "3" reporting conventional significance levels has found crime prevention effects for the program condition, and where the effect size is available, the effects are at least one-tenth of

one standard deviation better than the effects for the control condition OR the preponderance of evidence favors the program.

Don't Know (4): Categories with empirical evidence which do not fit one of the above are included in this residual category.
We must, of course, be extremely careful in labeling any program category as highly certain to be good or bad in its effects. Yet we must also be clear enough to make our conclusions useful, no matter how much we anticipate that science is always provisional and that our conclusions may be changed by next year. The large number of programs to be reviewed almost guarantees that some have strong evidence of extreme effects, both positive and negative. Yet most extreme results are from single, unreplicated studies. It is just as important not to conclude too much from a single negative result as from a single positive one. A single evaluation, even with strong evidence, cannot be assumed to generalize to all or most other settings. The primary objective of identifying promising results should be to foster replications, and the more promising the results the greater the replication need. Where there is substantial consistency of evidence in one direction, the senior author of each chapter makes a judgement and defends it in the text.

Here again we have *no* advice about question 2 – when is an evidence claim relevant to assessing the probability of effectiveness for a proposed policy; and faulty advice about question 3 – how to settle on the probability of effectiveness: to wit, don't look at the total picture. (This despite the fact that it is contrary to what we do in physics.)

So, we have arrived at my first point. We put a huge amount of effort and expense into question 1, which ensures that there really is a *causal* connection between programme and outcome somewhere. We do so in aid of evidence-based policy. But question 1 is not a policy question; it is a question in pure science. This is like Galileo rolling balls down an inclined plane or dropping them from the leaning tower of Pisa to test the effects of gravity on them. Galileo's experiments yield a great deal more precise information than any RCT; yet Galileo's results are a long way from the policy question: predicting the trajectory of our cannonballs under the influence of gravity.

Question 1 is a question in pure science; the policy-relevant issues about evidence are encoded in questions 2 and 3. Yet all the rigor, and almost all the attention, is to question 1. We are urged to extreme rigor at one stage, then left to wing it for the rest. But: a chain of defence for the effectiveness of a policy, like a towing chain, is only as strong as its weakest link. So the investment in rigor for one link while the others are left to chance is apt to be a huge waste. So if we do want decision-makers to use evidence as the basis for judgements about whether a policy will be effective when implemented, we need to develop far better guidelines for questions 2 and 3. Whatever else is going on, I am not surprised that evidence-based policies are not proving as effective as hoped. For if policy makers were assiduously following the advice of the current dominant guides, they would be making wrong assessments a great deal of the time.

How to use evidence: the need for a causal model

My second point is that the right answers to these two crucial questions in any particular case depend on the right choice of a causal model for that case, and advice on the two questions should reflect that.

Consider a simple case using every day physics. I choose this because it is simple, well understood and I am not likely to get involved in subject-specific debates in criminology. I have access to a desk magnet, alternatively to a large industrial magnet. I know the exact strengths of these with a very high degree of certainty – claims about their efficacy for lifting objects have passed far more than two good RCTs: they have centuries of study behind them. Shall I use one of them to lift an object in my driveway? That depends at this stage *entirely on features of the target situation.*

First, magnets need helping factors to be effective at all. My desk magnet is useless for lifting a matchstick; it is only the *combination* of a magnet and a metal object that produces a magnetic force. This has easy analogues in crime reduction. Consider the nice example of Nick Tilley and Ray Pawson. If CCTV cameras in car parks reduce car theft by discouraging thieves, they need to be visible to be effective at all. But if they work by alerting the police to get there in time to arrest the thieves, they had better be hidden. We need to know the necessary auxiliary factors.

Then the acceleration caused by the magnet is only one part of the story, often one very small part. To know what happens when we apply the magnet we need to know the other forces as well. Here, especially gravity. The desk magnet may lift a pin but it is hopeless for my car, where we need the industrial magnet. We also need to tend to what other forces we introduce in the course of getting the magnet in place. Perhaps the industrial magnet would have lifted the car if only we hadn't thrown the heavy packing case for the magnet into the boot. Finally, we need to know how all these factors combine to produce a result. Often in criminological or other social contexts we assume simple additivity: add a good thing and the results can only get better. But that doesn't work in even this simple physical case. We get so used to vector addition that we forget that it isn't simple addition of effect sizes. Add a magnetic acceleration of 42 ft/sec/sec to that of gravity's 32 ft/sec/sec and you won't necessarily get 74.

The point is that whether the magnet will be effective at all in the target situation and to what extent depends on the causal structure of the situation. So the most direct way of predicting its effects is to construct a causal model of the situation and estimate them.

I know no-one wants to hear this since it seems difficult. But consider: we know industrial magnets would pass any number of RCTs, of any degree of stringency. But that's not anywhere near enough to know. None of us would rent an industrial magnet

to remove a load of rubbish without looking at the *rubbish*. Knowledge that magnets just like this *can* lift is only a small part of what we consider when we evaluate whether renting the industrial magnet will be effective in removing our rubbish. If this is so in everyday calculations and in applied science and engineering, why should we expect it to be substantially different – and substantially easier – in social engineering

Of course constructing causal models is hard, even if the models are rough and we have figured out ways to tolerate the uncertainties. Sometimes there are shortcuts, "cheap heuristics". For instance, one powerful cause can swamp everything else so you don't need to model the rest. If you are going to put a bullet through someone's heart you do not need to find out what his cholesterol levels are to calculate his longevity. Or, as with the magnet and the matchstick, the absence of some necessary auxiliary can show that a policy will not be effective without further thought. For instance an elaborate schedule of rewards and punishments is not going to work in cases where people's actions aren't responsive to their utilities.

Failing a nice heuristic for a case, the right advice is: do your best with the resources and time available to build a causal model. This is my second point. We may not wish to build a causal model. We may not know how to; we may think it takes too much time or money, intelligence or attention. That does not alter the fact that when we buy a policy we are betting on a causal model, willy-nilly, whether we wish to think about it or not. Generally then it is better to think about it than not, and to do so in a systematic and deliberate way. Finally, tying the two points together: If what we are aiming for are reasonable causal models for our policy decisions, this provides direction for constructing advice for how to answer the three central questions, including the first, which I have ignored – the question of credible evidence claims. Because now we see we need information far beyond the kinds of causal claims that are the subject of the standard evidence- ranking systems. Those causal claims bear on one piece of the causal story. The guides show us how to decide if a magnet can lift – because it has definitely lifted somewhere in some circumstances. That does not tell us at all that it will lift in our circumstances. For that we need to know what the rubbish is like as well, what situation it is in and how all these factors behave together. These judgements should have solid backing as well if predictions are to be relied on.

To repeat, our assessment of the probability of effectiveness is only as secure as the weakest link in our chain of reasoning to arrive at that probability. We may have to ignore some issues or make heroic assumptions about them. But that should dramatically weaken our degree of confidence in our final assessment. Rigor isn't contagious from link to link. If you want a relatively secure conclusion coming out, you'd better be careful that each premise is secure going in.

SECTION II

EVIDENCE ON THE GROUND: WHAT RCTs CAN SUPPORT

2.1 A PHILOSOPHER'S VIEW OF THE LONG ROAD FROM RCTs TO EFFECTIVENESS

(From *The Lancet*, 2011)

For evidence-based practice and policy "RCTs are the gold standard." But exactly why? We know that RCTs do not, without a series of strong assumptions, warrant predictions about what happens in practice. But just what are these assumptions? I maintain that answers to both questions are obscured because we don't attend to what causal claims say. Causal claims entering evidence-based medicine (EBM) at different points say different things and failure to attend to these differences makes much current guidance about evidence for medical and social policy simplistic and misleading.

What a claim says and how it is warranted must slot together as in a jigsaw puzzle. The special virtues of RCTs are then a clue to the real content of the causal claims they warrant, and vice versa. So here I shall examine the evidential credentials of RCTs and from them derive what kind of causal claims they can support. Next I shall describe three different kinds of causal claims that commonly get conflated. Finally I shall argue that these three kinds of claims play very different roles in supporting effectiveness predictions and that the three kinds of claims need very different kinds of evidence to support them. The result is that we need a far more varied palette of kinds of evidence for predicting effectiveness than most prominent advice guides outline.

The first big question we need to be clear about is "What's so good about RCTs?" The canonical answer: "RCTs control for unknown confounders." This answer jumps into the middle of a discussion long underway. Two special features of ideally-conducted RCTs provide more fundamental grounding:

1. Ideal RCTs can *clinch* causal conclusions.
2. Ideal RCTs are *self-validating*.

Clinching. Some methods merely *vouch for* their conclusions. Though it is problematic to say exactly what it takes for a finding to vouch for a hypothesis, it generally involves at least that the finding is surprising but not surprising given the hypothesis. Others methods in the ideal *clinch* their conclusion: If the assumptions defining the method are met, positive results *deductively imply* the conclusion. The ideal RCT – ie, one for which all the requisite premises are met – is a clincher. Roughly, RCT logic assumes a general metaphysical premise (1): Probabilistic dependence calls for causal explanation. Experimental design acts to ensure premise (2): All features causally relevant to the outcome other than the treatment (and its downstream effects) are distributed identically between treatment and control groups. If [premise (3)] the outcome is more probable in the treatment than the control group, the only explanation possible is that the treatment caused the outcome in some members of that group.

EBM focuses on clinchers. Yet studies in philosophy of science suggest that physics claims are primarily warranted by large collections of varied results merely vouching for them. There are no checklists for handling vouching evidence, however; perhaps this is why EBM guidelines favour clinchers. Clinching is not unique to RCTs however.

Economists use the rigorous methods of econometric modelling to estimate the degree to which one factor predicts another in a given population. This could be mere correlation. But given the right assumptions their results can deductively imply causal conclusions from non-experimental data. Deduction from accepted theory can also clinch causal conclusions; as can ideal case-control studies, since these have the same logic as RCTs. The difference between RCTs and these others is the grounds for accepting the requisite premises, which is the second special feature of RCTs.

Self-validating. All methods have assumptions that must be met before conclusions from them are warranted. For causal conclusions, some of these premises must be causal: "No causes in; no cause out." For most studies – eg, the economic ones mentioned – the warrant for these assumptions comes from outside the study design.

The metaphysical assumption aside, support – though no guarantee – for premises 2 and 3 is built right into RCT design. Premise 2, by policing of treatment administration, blinding, random assignment, etc.; premise 3, by techniques – including large sample size – for reliably inferring probabilities from observed frequencies. RCTs are thus *self-validating*.

Self-validation is a virtue but not a necessity. We often have good reason to accept the premises necessary for other study designs, including case-control studies, which is where unknown confounders enter. By definition we do not know "unknown" causal factors. We may nevertheless know enough about underlying mechanisms and/or the study environment to assume no strong unknown causes obtain. Sometimes we even shield studies to prevent unknown sources of confounding, eg, by conducting magnetic-resonance studies in Hertz boxes. RCTs trust to procedure; other methods import information. Which strategy provides most support for a particular conclusion depends on how confident we can be that the procedures achieve their aim in the case at hand versus the strength of justification for the information imported.

All the studies I have discussed so far justify "efficacy claims", where "efficacy" is what happens in *ideal* circumstances. But recall the logic of RCTs. The circumstances there are ideal for ensuring "The treatment caused the outcome in some members of the study"; ie, they are ideal for supporting "*it-works-somewhere*" *claims*. But they are in no way ideal for other purposes; in particular they provide no better base for extrapolating or generalizing than knowledge that the treatment caused the outcome in any other individuals in any other circumstances (except in the very unusual situation where there is good reason to think that the study population is a representative sample of the target population).

For policy and practice we do not need to know "it works somewhere". We need evidence for "*it-will-work-for us*" *claims*: The treatment will produce the desired outcome in our situation as implemented there.

How can we get from it-works-somewhere to it-will-work-for-us? Perhaps by simple enumerative induction: Swan 1 is white; swan 2 is white; So the next swan will be white. For this we need a large and varied inductive base – lots of swans from lots of places; lots of RCTs from different populations – plus reason to believe the observations are projectable, plus an account of the range across which they project. Electron charge is projectable everywhere – one good experiment is enough to generalize to all electrons; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

Sometimes it can. Magnets are tested in ideal circumstances; their power to attract metal objects can be relied on widely. The Heimlich manoeuvre is good for removing airway obstructions in almost anyone and aspirins are generally a good bet for relieving headaches. Knowledge like this involves a third kind of causal claim, a *power* or *capacity claim*: The treatment *reliably promotes* the outcome, or reliably contributes across a given range of circumstances. "Reliably promotes" means roughly that across a wide range of circumstances there will be more cases, or a higher level, of the outcome with the treatment than there would be without it. What the actual numbers are depends on what other factors are present, just as the actual motion of a pin attracted by a magnet depends on gravity, the wind, etc.

Where available, knowledge of capacities is a powerful tool. To use RCT results as evidence for effectiveness we are generally told to look for populations/settings like those of the study. This is advice difficult to follow. We do RCTs because we do not know all the

major relevant factors, so judging whether other situations are relevantly similar is hard. Moreover, similarity is rare. But then, similarity is not necessary if the treatment reliably promotes the outcome. Magnets attract metal objects almost everywhere. The Heimlich manoeuvre depends on almost universally shared structures in the human body, so it can be relied on to encourage removal of obstructions across a wide variety of settings and individuals. So capacity claims provide evidence for effectiveness even in situations very different from those of any study. And where no capacity claims obtain, there is seldom warrant for assuming that a treatment that works somewhere will work anywhere else. (The exception is the one noted, where there is warrant to believe that the study population is a representative sample of the target population – and cases like this are hard to come by.)

But there are problems for using capacity claims. First, although knowledge that a treatment reliably promotes an outcome is evidence that it will cause that outcome for us, it is only *part* of an evidential argument. We also need to know that our situation contains all requisite helping factors and that there are no overwhelming countering causes. Magnets lift objects only if the objects are metal and they will not lift even metal objects when gravity is too strong. Nor will the Heimlich manoeuvre remove objects if the oesophagus is too swollen by disease; many powerful medicines will not work if certain items are missing from the diet; and homework is generally an aid to learning only given a quiet, supportive environment in which to do it. I highlight these additional factors not because they are unfamiliar but because influential guidelines for evidence-based medical and social policy often do not mention them let alone discuss standards of evidence for claims about them – despite the fact that such information is necessary for any reasonable predictions about effectiveness.

Second, capacity claims are hard to warrant. Worse, there is no explicit methodology describing exactly what it takes to warrant them, even in physics, despite the fact that most of our successful interventions using physics depend on capacity claims. What is clear is that even a handful of RCTs by themselves will not do the job. In general to support a capacity claim, a general understanding is needed of *why* the treatment should have the power to produce the outcome. Happily this is often available though few guidelines direct us to look out for it, let alone provide advice about what counts as good evidence that the backup understanding is sound. Probably that's because we try to rely on procedures, as with RCTs, to avoid relying on claims of a general theoretical nature.

But an RCT supports only an "it-works-somewhere" claim. How can we put hard-won RCT results to use for predicting "it will work for us"? Similarity is problematic to judge and the kind of similarity necessary for warranting direct extrapolation from RCTs is rare. Capacities provide a conduit from RCTs to effectiveness, often the only one. But these are hard to warrant and even when warranted are only part of a good evidence base for predicting effectiveness. Effectiveness predictions are always dicey. Use of scientific evidence makes them far less so. But to use this evidence we need to tackle, not ignore, the messy issue of "theoretical" warrant for capacities in medical and social contexts.

Further reading

Cartwright N.D. (2009) 'What is this Thing Called "Efficacy"?' in *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice*, C. Mantzavinos (ed), Cambridge: Cambridge University Press. Available at: http://personal.lse.ac.uk/cartwrig/Default.htm

Reiss, J. (2010) 'Empirical Evidence: Its Nature and Sources', in Jarvie, Ian C. and Jesus Zamora Bonilla (eds), The Sage Handbook of Philosophy of Social Science, Thousand Oaks (CA): SAGE Publications. Available at: www.jreiss.org

Worrall, J.: (forthcoming) 'Evidence: Philosophy of Science meets Medicine', *The Journal of Evaluation in Clinical Practice*.

——— (2007) 'Evidence in Medicine and Evidence-Based Medicine', *Philosophy Compass* 2/6 981-1022.

2.2 ARE RCTs THE GOLD STANDARD?

(From *BioSocieties,* 2007*)*

The answer to the title question, I shall argue, is "no". There is no gold standard; no universally best method. Gold standard methods are whatever methods will provide (a) the information you need, (b) reliably, (c) from what you can do and from what you can know on the occasion. Often randomized controlled trials (RCTs) are very bad at this and other methods very good. What method best provides the information you want reliably will differ from case to case, depending primarily on what you already know or can come to know. Since I have no expertise in psychiatry, I shall discuss methods in general use in the human sciences without trying to approach special problems of psychiatry. The article will have six parts:

- Clinchers vs vouchers: a distinction and its implications
- A straddler: the hypothetico-deductive method
- Examples of methods that clinch conclusions
- RCTs: ideal RCTs, real RCTs and the scope of an RCT
- The vanity of rigor in RCTs
- Closing remarks

Bits of the section "RCTs: ideal RCTs, real RCTs and the scope of an RCT" will rely on some formal results that I will present informally. I hope to convey a sense of the kind of information that is required to justify the claims of RCTs to be a gold standard, both as a basis for caution and as a basis for comparison with other methods that have an equal claim to this status (because they are what I shall call "clinchers").

Clinchers vs vouchers: a distinction and its implications

Methods for warranting causal claims fall into two broad categories:

1. Those that *clinch* the conclusion but are *narrow* in their range of application, for example RCTs, derivation from theory or certain econometric methods.

2. Those that merely *vouch for* the conclusion but are *broad* in their range of application, for example qualitative comparative analysis, or looking for quantity and variety of evidence.

What is characteristic of methods in the first category is that they are deductive: *if* all the assumptions for their correct application are met, then if evidence claims of the appropriate form are true, so too will the conclusions be true. But these methods are concomitantly narrow in scope. The assumptions necessary for their successful application will have to be extremely restrictive and they can take only a very specialized type of evidence as input and special forms of conclusion as output. That is because it takes strong premises to deduce interesting conclusions and strong premises tend not to be widely true. Methods in the second category are more wide-ranging but it cannot be proved that the conclusion is assured by the evidence, either because the method cannot be laid out in a way that lends itself to such a proof or because, by lights of the method itself, the evidence is symptomatic of the conclusion but not sufficient for it. What then is it to *vouch for*? That is hard to say since the relation between evidence and conclusion in these cases is not deductive and there are no general good practicable "logics" of non-deductive confirmation, especially ones that make sense for the great variety of methods we use to provide warrant.

The fact that RCTs are a deductive method underwrites their claims to be the gold standard. But RCTs suffer, as do all deductive methods, from narrowness of scope. Their results are formally valid for the group enrolled in the study, but only for that group. The method itself does not underwrite any strong claims for external validity, that is for extending whatever results are supposed to be established in the test population to other "target" populations. This is important to keep in clear sight in comparing RCTs with other methods.

Compare, then, the costs and benefits of the two categories. Clinchers are deductive: *if* they are correctly applied *and* their assumptions are met, then *if* our evidence claims are true, so too will be our conclusions – a huge benefit. But there is an equally huge cost. These methods are concomitantly narrow in scope. The assumptions necessary for their successful application (a) tend to be extremely restrictive, (b) can only take a very specialized type of evidence as input, and (c) have only special forms of conclusion as output. In consequence we face a familiar kind of trade-off: we can ask for methods that clinch their conclusions but the conclusions are likely to be very limited in their range of application.

A straddler:
the hypothetico-deductive method

The hypothetico-deductive method is a straddler. Used one way – the way Karl Popper advocated – it is purely deductive and so is in the same category as the RCT. The method works, as all methods do, by presupposing a variety of auxiliary assumptions, otherwise nothing really follows from the hypothesis of interest.

Popper:
Hypothesis → outcome
¬outcome
Therefore, ¬hypothesis

This is a clincher.

Positivists:
Hypothesis → outcome
outcome
probability of the hypothesis increases
(ceteris paribus)

This is a voucher.

Popper argued that the only correct use of the hypothetico-deductive method is as a clincher, to deduce that hypotheses are false. The argument accepted by the Positivists, he pointed out, is a deductive fallacy – the fallacy of affirming the consequent. And deductive logic, he maintained, is

all the logic there is. This is borne out by centuries of failed efforts to establish some reasonable relatively uncontroversial theory of inductive confirmation. On the other hand, philosophers of physics maintain that the hypothetico-deductive method is the method by which physics theories are established. Nevertheless, medical science – and most of current evidence-based policy rhetoric – will not allow it.

Perhaps an example related to topics of interest to psychiatry will help. Consider the widespread correlation between low economic status and poor health, and look at two opposing accounts of how it arises (for a discussion and references see Cartwright, 2007). Epidemiologist Michael Marmot from University College London argues that the causal story looks like this:

Marmot:
Low status → "stess" → too much "fight or flight response → poor health

In contrast, Princeton University economist Angus Deaton suggests this:

Deaton:
Poor health → loss of work → low income → low status

Deaton confirms his hypothesis in the National Longitudinal Mortality Study (NLMS) data. He reasons: if the income–mortality correlation is due primarily to loss of income from poor health, then it should weaken dramatically in the retired population where health will not affect income. It should also be weaker among women than men, because the former have weaker attachment to the labour force over this period. In both cases these predictions are borne out by the data. Even more, split the data between diseases that something can be done about and those that nothing can be done about. Then income is correlated with mortality from both – just as it would be if causality runs from health to income. Also, education is weaker or uncorrelated for the ones that nothing can be done about. Deaton argues that it is hard to see how this would follow if income and education are both markers for a single concept of socio-economic status that is causal for health.

Thus Deaton's hypothesis implies a number of specific results that are borne out in NLMS data and would not be expected on dominant alternative hypotheses. So the hypothesis seems to receive positive confirmation, at least if we share the Positivists' intuition. More carefully, it seems to receive some confirmation for the population sampled for the NLMS data. But what about other populations, ie, what about *external validity*? The arguments I have just described that seem, contra Popper, to provide some evidence for Deaton's hypothesis in the population sampled, do nothing as they stand to support any claims about alternative populations. More premises and more and different arguments are needed to do that. So here we are reminded how badly even a non-clinching method can suffer from problems of external validity.

Examples of methods that clinch conclusions

I list just a few other kinds of methods that work deductively.

- Econometric methods
- Galilean experiments
- Probabilistic/Granger causality
- Derivation from established theory
- Tracing the causal process
- Ideal RCTs

These are clinchers: it can be proved that if the auxiliary assumptions are true, the methods are applied correctly and the outcomes are true and have the right form, then the hypothesis must be true. Even though I do not have the space to discuss them here, I mention them in order to stress that, when it comes to clinchers – to methods from which the hypothesis can be rigorously derived from the evidence – RCTs are not the only game in town. There are lots of methods that can clinch conclusions.

It is important to keep in mind one caution, however. To buy the benefits of a clinching method we must be able to ensure that it is highly probable that *all* the requisite premises are obtained. That's because of the *weakest link* principle for deductive reasoning. The probability of the conclusion can be no higher than that of the weakest premise.

– Suppose you have 10 premises, 9 of them almost certain, one dicey. Your conclusion is highly insecure, not 90 per cent probable.
– In a deductive argument P(conclusion) ≤ P(conjunction of premises)

I belabour this because of the benefits of clinching methods – clinchers are rigorous. It is transparent *why* the results are evidence: given the background assumptions the hypothesis follows deductively from the results. And it is transparent *when* the results are evidence: when the background assumptions are met. This contrasts with ethnographic methods and expert judgment, for example. These can provide extremely reliable evidence. But there is no specific non-trivial list of assumptions that tell when they have done so. But if you want credit for this benefit of a clinching method, you must be able to show that the *conjunction* of your premises has high probability *in the case at hand*.

Randomized controlled trials

Ideal RCTs

I have claimed that ideal RCTs are clinchers. That of course depends on how they are defined. But there are perfectly natural definitions from which it can be proved that RCTs, as thus defined, allow causal claims about the population in the study to be deduced from probability differences between the treatment and control groups (cf. Cartwright,1989; Heckman, 2001; Holland and Rubin, 1988). The one I have worked with extensively is the probabilistic theory of causality, formalized by Patrick Suppes (1970) but widely adopted throughout the human sciences, even if not consciously so under that title. Suppes' concept of probabilistic causality is similar to the concept of Granger causality (Granger,1969) that is frequently used in econometrics.

The root idea of the probabilistic theory of causality is that if the probability of an "outcome" O is greater with a putative cause T than without T, once all "confounders" are controlled for in some particular way, that is sufficient for the claim "T causes O" in that particular setting of confounding factors. So, in a population where "all other" causes of O are held fixed, any difference in probability of O with T present versus with T absent shows that T causes O in that population. The rationale supposes that differences in probability need a causal explanation, and, if all explanations relying on confounders are eliminated, then T causes O is the only explanation left. T must be causing O in at least some members of the population in order to account for the difference in probability. I should note that whether one wishes to adopt the theory in exactly this form, some such assumption is necessary to connect causes and probabilities if we are to suppose that the probabilistic observations in RCTs can yield causal conclusions.

The definition so far only tells us when we can assert that T causes O for populations that have some fixed arrangement of "all other" causal factors. To get a more general conclusion we may accept as well that if T causes O in a subpopulation of a given population ϕ, then T causes O in ϕ. This is consistent with my suggestion in the last paragraph that, on the probabilistic theory of causality, when we say T causes O in a population we mean that T causes O in at least some members of that population.

The proof that positive results in an ideal RCT deductively imply that the treatment causes the outcome would go something like this: to test "T causes O" in ϕ via an RCT, we suppose that we study a test population ϕ all of whose members are governed by the same causal structure (CS), for O and which is described by a probability distribution P. P is defined over the event space $\{O,T,K_1,K_2,\ldots,K_n\}$, where each K_i is a state description over "all other" causes of O except T.[1] The Ki are thus maximally causally homogeneous subpopulations of ϕ. Roughly:

– "K_i is a state description over other causes" = K_i *holds fixed* all causes of O other than T.
– "Causal structure" = the network of causal pathways by which O can be produced, with their related strengths of efficacy.

Then assume
1. *Probabilistic theory of causality*. T causes O in ϕ if P(O/T&Ki) > P(O/¬T&Ki) for some subpopulation Ki with P(Ki) > 0.
2. *Idealization.* In an *ideal* RCT for "T causes O in ϕ", the Ki are distributed identically between the treatment and control groups.
From 1 and 2 it follows that ideal RCTs are clinchers. If P(O) in treatment group > P(O) in the control group in an ideal RCT, then trivially by probability theory P(O/T&Ki) > P(O/ ¬T&Ki) for some Ki. Therefore: if P(O) in treatment group > P(O) in control group, T causes O in ϕ relative to CS,P.

What is going on here? We suppose that increase in probability of O with T does not show that T causes O in an arbitrary population. But it does in a maximally causally homogeneous population. We of course are almost never in a position to identify what makes for a maximally homogeneous population, so how can we tell whether T increases the probability of O in some one of these? The RCT is a clever way to find out. The RCT tells us that in some one or another maximally causally homogeneous subpopulation of the population in the study, T does increase the probability of O. Given the probabilistic theory of causality that tells us that T causes O in that subpopulation. So, what is established in the ideal RCT, according to the account based on probabilistic theory of causality, is that T causes O in at least one maximally causally homogeneous subpopulation of ϕ. We may say we have established "T causes O in ϕ" and that is a fine way to talk, so long as we recall that this means that T causes O in some subpopulation of ϕ.

It is important to notice that, on this account, "T causes O in ϕ" is consistent with "T causes ¬O in ϕ". This lines up with what we know of RCTs:

– RCTs deliver population-average results. A *positive* result shows that T causes O in at least one subpopulation. It could produce exactly opposite results in other subpopulations.

– Positive results are conclusive but negative ones are not. Equal probability for O in the treatment and control groups does not show that T does not cause O in ϕ. It shows that if T causes O in ϕ (because it does so in some Ki $\subseteq$ ϕ) it must also cause ¬O (because it does so in some other Ki $\subseteq$ ϕ).

Real RCTs

So, from positive results in an ideal RCT for "T causes O in ϕ", we can deduce that the causal hypothesis is true. But we can be no more certain of our casual conclusion than we are of our premises, to wit, that the RCT is ideal and that the probability of O is indeed higher with T than without in the test population. What do we do to ensure the premises? Here are just some of the principal precautions we take: careful use of statistics to move from frequencies to probabilities, "random" assignment to treatment and control groups, quadruple blinding, careful attention to drop-outs and non-compliance, and so on.
I mention them just to point out that the practical methodology must match and be matched with the kind of formal treatment I have outlined. RCT advocates claim that RCTs are extremely reliable if carried out properly. That claim can be justified by an account of the kind I have outlined. But then—what is justified is that positive results *as defined by the account*, in an ideal RCT *as defined by the account*, imply a causal conclusion *of the kind defined by the account*. The practical

[1] This must include 'spontaneous generation'. More formally, Ki holds fixed one variable on each pathway that does not go through T, as judged by the causal structure CS.

methodology then must be geared to ensuring that the premises required by the formal account are very likely to be true; and the conclusions drawn can only be of the kind admitted by the account. Of course the converse holds as well: a formal account that does not match well with our most careful, most well thought-out practical methodology should be viewed with at least a little suspicion.

The Scope of an RCT

Starting as I have from the probabilistic theory of causality, there are two kinds of causal conclusions we might naturally try to export from an RCT to some target population θ:

1. T causes O in θ. That is, T causes O in at least some members of θ.[2]
2. Some measure of "average improvement" that holds in the experiment will hold in the target population. I shall consider the simple case of P(O/T) > P(O/¬T).

Both conclusions need strong auxiliary assumptions to be warranted, well beyond those supported by the structure of the RCT. For the first, the RCT shows that T causes O in at least some members of some fixed causally homogeneous subpopulations. So to draw conclusions that T causes O in at least some members of θ, we need at least these kinds of assumptions:

– Auxiliary 1.a. At least one of the subpopulations (with its particular fixed arrangement of "other" causal factors) in which T causes O in ϕ is a subpopulation of θ.

– Auxiliary 1.b. The causal structure and the probability measure is the same in that subpopulation of θ as it is in that subpopulation of ϕ.

For the second kind of conclusion we need to show that the outcome is more probable with T than without in θ.[3] The simplest guarantee for this is

– Auxiliary 2. The causal structure (CS) and the probability (P) are the same in θ as in ϕ.

There is an indefinite number of other ways that guarantee P(O/T)>P(O/¬T) in θ given it holds in ϕ, depending on the exact strengths of efficacy and the exact probabilities involved. But this is the only rule that does not require explicit statement of the specific numbers, most (if not all) of which are unknown to us. To get a sense for this, just imagine a case where there are only two relevant subpopulations, in one of which T is strongly positive for O and in the other it is equally strongly negative. The results will be positive in the RCT if the first subpopulation is more probable than the second, but will be reversed in targets where the second outweighs the first even

[2] In the 'long run', of course, since all results are probabilistic.

[3] Or as near enough as matters for our purposes. I shall here ignore these niceties and how to treat them in order to focus on the main point.

if the new population has the same causal structure as the test population. Clearly if the causal structure differs, matters will depend on just how, just as the net result will depend on just what the probabilities are if the probabilities of the relevant subpopulations differ.

The central question for external validity then is, "How do we come to be justified in the assumptions required for exporting a causal claim from the experimental to a target population?" Here rigor gives out. This is not to say that we do not have procedures or that we do not proceed in an intelligent way. We could aim to draw the test population "randomly" from the target. We know that this is almost never possible. Moreover, we must not be deluded about sampling methods: You cannot sample randomly without any idea what factors are to be equally represented – which is just the issue that drives us to RCTs to begin with. One thing we certainly can do is to try to take into account all possible sources of difference between the test and target populations that we can identify. This is just what we do in matched observational studies. When it comes to internal validity, however, advocates of the exclusive use of RCTs do not take this to be good enough – matching studies are not allowed just because our judgements about possible sources of difference are fallible. Yet exactly the same kinds of "non-rigorous" judgments are required if RCTs are to have any bearing outside the test population. For an RCT, the reliability of the claims in the target population is only as good as our estimates that very demanding auxiliaries like those above are met. The question then is about the trade-off between internal and external validity.

Lesson. We experiment on a population of individuals each of whom we take to be described (or "governed") by the same *fixed causal structure* (albeit unknown) and *fixed probability measure* (albeit unknown). Our deductive conclusions depend on that very causal structure and probability. How do we know what individuals beyond those in our experiment this applies to? We have seen some typical auxiliary assumptions about target populations that allow us to export conclusions from the experimental population to a target population, and we have seen that these assumptions are very demanding – demanding of information that is not supplied by the RCT and that is hard to come by. But our conclusions about the target can be no more certain than these auxiliary assumptions. The RCT, with its vaunted rigor, takes us only a very small part of the way we need to go for practical knowledge. This is what disposes me to warn about the vanity of rigor in RCTs.

The Vanity of Rigor in RCTs

The title is borrowed from my article "The vanity of rigor in economic models" (in Cartwright, 2007). In both cases we see identical problems: that of internal versus external validity. Economists make a huge investment to achieve rigor inside their models, that is, to achieve internal validity. But how do they decide what lessons to draw about target situations outside from conclusions rigorously derived inside the model? That is, how do they establish external validity? We find: thought,

discussion, debate; relatively secure knowledge; past practice; good bets. But not rules, check lists, detailed practicable procedures; nothing with the rigor demanded inside the models.

And RCTs? If we compare them with economic models on internal validity, economic models have the advantage: We can readily see when the results are internally valid in an economic model just by inspecting the derivation. This is clearly not so with RCTs. Consider the equal distribution of "other" causal factors. Once we check the causes we know about, we have no further evidence that our precautions, our quadruple blinding and random assignment and so forth, indeed result in an equal enough distribution. And we know lots of things can go wrong. The best we can do is for people expert at what could go wrong to have a very close look at what actually happens in the experiment. It is important though that these are not people like me (or independent experimental- design firms) who know only about methodology, but rather people with subject-specific knowledge who can spot relevant differences that come up. But this introduces expert judgment into the assessment of internal validity, which RCT advocates tend to despise. Without expert judgment, however, the claims that the requisite assumptions for the RCT to be internally valid are met depend on fallible mechanical procedures. Expert judgments are naturally fallible too, but to rely on mechanics without experts to watch for where failures occur makes the entire proceeding unnecessarily dicey.

This brief mention of economic models versus RCTs highlights the conventional trade-off I recalled at the start between internal and external validity. Despite the claims of RCTs to be the gold standard, economic models have all the advantages when it comes to internal validity. As I remarked, we need just mathematics and logic to decide if the conclusions are internally valid, whereas RCTs need a number of demanding assumptions beyond valid reasoning. But it seems that RCTs have the advantage over economic models with respect to external validity. Surely, no matter what the target population, people in experiments are more like people in the target population than people in models are. Even here there is a caution, however, for of course this claim depends on exactly what kind of knowledge about people in the target population we build into the construction of our experiments versus how much we build into our models, and how we do so.

Closing Remarks

I close with some reminders for those who advocate RCTs as the gold standard. The method of our most successful science – the h-d method – is not a clincher at all. (And we do have some biomedical theory!)

There are many other clinching methods. Which method provides the most secure conclusions in a given case depends entirely upon which kinds of premises we can be most secure about and the situation at hand.

An argument that certain procedures achieve a given result much of the time may not be a good argument that they do so on any one occasion. External validity for RCTs is hard to justify. Other methods, less rigorous at the front end, on internal validity, can have far better warrant at the back end, on external validity. We must be careful about the trade-offs. There is no a priori reason to favour a method that is rigorous part of the way and very iffy thereafter over one that reverses the order or one that is less rigorous but fairly well reasoned throughout.

References

Cartwright, N. (1989). Nature's capacities and their measurement. Oxford: Clarendon Press.

——— (2007). *Hunting causes and using them.* Cambridge: Cambridge University Press.

Granger, C. (1969). Investigating causal relations by econometric models and cross-special methods. *Econometrica, 37,* 424-438.

Heckman, J. (2001). Econometrics, counterfactuals and causal models. Keynote Address, International Statistical Institute, Seoul, Korea.

Holland, P.W., & Rubin, D.B. (1988). Causal Inference in retrospective studies. *Evaluation Review, 12,* 203-231.

Suppes, P. (1970). *Probabilistic theory of causality*. Atlantic Highlands, NJ: Humanities Press.

2.3 RCTs, EVIDENCE AND PREDICTING POLICY EFFECTIVENESS

(From *The Oxford Handbook of the Philosophy of the Social Sciences,* 2012*)*

1. Introduction

"To Test Housing Program, Some Are Denied Aid", so says a headline in the *New York Times* published December 8, 2010. Why were they denied aid? Because they were in the control wing of a randomized controlled experiment (RCT) to determine for a New York City Homeless Services program called Homebase "whether the $23 million program… helped the people for whom it was intended".

Apparently New York City has bought into a standard claim in the movement for evidence-based policy and practice (EBPP): that RCTs are the gold standard for establishing "what works". So too it seems has the Greater London Authority whose Project Oracle "aims to establish a coordinated London wide way of understanding and sharing what really works". That's according to the Authority's *Standards of Evidence* document, which later explains that "the Standards value evaluations that use a comparison or control group … If the children going into the comparison or control group are assigned at random, so much the better as far as confidence in the results is concerned."

And according to the same *New York Times* article so too does the US federal Department of Housing and Urban Development which is also doing a RCT on families in homeless shelters. The article also reports on the very vigorous movement to introduce RCTs widely in development economics, spearheaded by a group of MIT economists:

Such trials, while not new, are becoming especially popular in developing countries… "It's a very effective way to find out what works and what doesn't," said Esther Duflo, an economist at the Massachusetts Institute of Technology who has advanced the testing of social programs in the third world. "Everybody, every country, has a limited budget and wants to find out what programs are effective" (Buckley 2010).

I urge we resist this movement. RCTs, if done very well, can indeed establish something about a program – that it worked in the studied situation. Just that, according to the *New York Times* is what the Department of Housing and Urban Development is trying to do: "The goal, a HUD spokesman, Brian Sullivan, said, is to find out which approach most effectively ushered people into permanent homes." Pay careful attention to the form of the verbs. To show that a program usher*ed* people into homes – ie, that the program work*ed* for the population studied – is a long way from establishing that it *will work* in a particular target, let alone that it *works* in general. I will here show just how long – and tortuous – this way is. That however is not the impression you get from the Duflo quote nor from the GLA document, which propose to use RCTs to find out "what *works*".

I have for a while been urging that we need a theory of evidence in terms of which we can evaluate strong claims like these for RCTs (Cartwright 2009; Cartwright and Stegenga 2009, 2011). Here I shall provide such a theory, a simple straightforward theory that allows us to see just what must be in place for RCT evidence to be relevant to predicting that a program or treatment will work in a target setting. Often the claim that a program worked in a studied population is called an *efficacy* claim and the prediction that it will work in a target, an *effectiveness* claim. It is widely acknowledged that the two are not the same. But little is said about how to get from one to the other and often it seems to be implied that the fall back position is that it is a reasonable bet that a program that was efficacious somewhere will be effective in any other situation unless we have specific reason to think it won't work in that other situation, especially if the two have superficial similarities that are salient in the discipline proposing the inference (eg, demographic features like urban versus rural; physiological features like gender or age; or socioeconomic ones like class, wealth or religion). On the contrary, I shall argue, the demands that must be met for efficacy to count as evidence for effectiveness at all, let alone sufficient evidence, are high and need good reasons in their favor if this bet is to be reasonable.

Before turning to the "theory of evidence for effectiveness claims" I shall first explain what RCTs can show, and why.

2. RCTs and what can be inferred from them

I shall describe here an ideal RCT. Randomization of subjects to treatment and control groups, blinding and large study populations are supposed to provide warrant for the assumption that a real RCT approaches the ideal – or near enough. This assumption is controversial (Worrall 2002, 2007). But I should like to sidestep this controversy and consider what can be inferred from a positive RCT result if we are willing to take it as well-established.

An ideal RCT for cause X and outcome Y randomly assigns individual participants in the study, $\{u_i\}$, into two groups where X = x for some value x universally in one group (the treatment group) and X = x' ≠ x universally in the other (the control group). No differences are to obtain in the two groups other than X and its downstream effects.

The standard result measures the so-called "*treatment effect*", *T,* across the units participating in the study (letting $\langle\theta\rangle$ represent the expectation of θ):

2.1. T =df $\langle Y(u)/X(u) = x\rangle - \langle Y(u)/X(u) = x'\rangle$
Of what interest is this strange statistic about randomized units in a study group – and familiarity should not make you forget that it is strange? There are two standard answers. One relies on the counterfactual analysis of Paul Holland and Donald Rubin (Holland and Rubin 1988). This is for instance the method of analysis adopted by James Heckman, who won the Nobel Prize for his work evaluating social programs, especially on selection bias and particularly with respect to labor markets (Heckman 2001, 2005)[1]. Let $Y_x(u)$ =df the value u would have for Y if X were set at x by a David Lewis-type miracle (which changes only X and its downstream effects).

[1] See also my discussion of Heckman in Cartwright 2007.

Now consider the expectations for the difference in counterfactual values of Y across units in the study. That is, the difference in Y each individual would experience were they treated with x' versus with x, and take the average. So:

2.2. *<CD>:* $\langle Y_x(u) - Y_{x'}(u)\rangle = \langle Y_x(u)\rangle - \langle Y_{x'}(u)\rangle$.

Notice that this is not the same as T since we can't observe $\langle Y_x(u)\rangle$ across all the units in the experiment but only $\langle Y_x(u)/ X(u) = x\rangle$ or the same for X = x'. But we may suppose that randomization – in the ideal at least – guarantees that for u's in the study the value u would have for Y if u received the treatment or control is (probabilistically) independent of whether u gets the treatment or control. Then

$\langle Y_x(u)\rangle - \langle Y_{x'}(u)\rangle = \langle Y(u)/X(u) = x) - Y(u)/X(u) = x'\rangle$

Or

T = <CD>: the observed treatment effect = the expectation of the counterfactual difference.

What use is the treatment effect under this interpretation, that is, where it is taken to be the expectation of the counterfactual difference? It's surely good for post hoc assignment of responsibility. It tells us that X definitely contributed to the Y values of some individuals in the study and further tells us the average contribution. For these individuals, X is definitely to blame – or praise – for part of their Y values. This is *attribution* or *evaluation* – which is what, according to the *Times* article, the HUD spokesman said their RCT was aimed at. RCTs can be very good for that. But beyond that?

We can look for so-called "external validity". Where else does the same result hold?

Range of external validity = those situations where the same result holds.

But with the assumptions so far, I see no way to begin to answer this question. Here the canonical trade-off between internal validity and external looms large. The study is perfectly geared to yield the average counterfactual difference. But without a lot more assumptions – both substantive and metaphysical – there's no place to go with it.

Notice that the counterfactual interpretation makes no mention of causal principles. The other common analysis of RCTs postulates that Y values for the units in the study (or their probabilities) are determined by a causal principle. The RCT can tell us something about the role of X, if any, in this principle.

Suppose then that Y in the study population of individuals is determined by the causal principle L:

L: $Y(u) \mathrel{c}= \alpha(u) + \beta(u)X(u) + W(u)$[2]

[2] The symbol '$\mathrel{c}=$' implies that the left- and right-hand side are equal and that the factors on the right-hand side are causes of the one on the left.

where W represents the net contribution of causes that act additively in addition to X, and X may not play a role in the equation at all if $\beta = 0$. The formula makes clear that β not only determines whether X contributes to Y at all but it also controls how much a given value of X will contribute to Y.

From L and the definition of T it follows that

$T =_{df} \langle Y(u)/X(u)=x\rangle - \langle Y(u)/X(u)=x'\rangle$

$= \langle \alpha(u)/X(u) = x\rangle - \langle \alpha(u)/X(u) = x'\rangle + \langle \beta(u)/X(u) = x\rangle x - \langle \beta/X(u) = x'\rangle x' + \langle W(u)/X(u) = x\rangle - \langle W(u)/X(u) = x'\rangle$.

If we are prepared to suppose that the random assignment of units to x and x' assures that for units in the study, X is probabilistically independent of α,β,W, then

2.3. $T = \langle \beta(u)\rangle (x - x')$

If T is positive then β is too. So X genuinely appears as a cause for Y in the law L for the study population. If $\beta = 0$ for all units then X does not appear in L. So under L, X makes no contribution to Y outcomes; these are produced entirely by the quantities represented in the variable W.

Note that the effect size of X with respect to Y does not tell the actual value of Y that occurs, nor its mean; rather it tells only the *contribution* of X. What actually happens, or happens on average, depends on W as well. And this can be a problem. Sometimes things are getting worse naturally (due to the action of factors in W); then the net results after the policy may be worse than before even though the policy improved matters over what they otherwise would have been. Sometimes we make matters worse ourselves in implementing the policy, as in the California class-size reduction program, where class-sizes were reduced but so too, due to the sudden need for many more teachers, was teacher quality (Bohrnstedt and Stecher 2002). So what's in W matters to forecasting actual results. Here though I shall lay aside consideration of factors in W in order to concentrate on the effect size and what can be learned from it.

3. What we want for predicting effectiveness and what's on offer

For evidence-based policy we want to predict with reasonable confidence[3] that the proposed policy will contribute positively to targeted outcomes in our situation as the policy would in fact be implemented there. What characterizes a body of evidence that can do this for us? I propose a broad swipe at a straightforward answer, an answer that works for any empirical hypothesis H. This proposal is clearly very crude but it will suffice for moving the discussion along:

[3] I don't imagine that we will generally be able to assign numerical probabilities to these hypotheses in any reasonable way; but we can certainly often make sound judgments about when a hypothesis is fairly well supported given the current state of knowledge and when badly supported. For policy prediction I would expect that most often the support that we can muster is weak. If so, we should acknowledge that and manage the uncertainty in sensible ways, not pretend we have assurances we lack.

H is well supported by a set S of empirical facts if

(A) The facts (fact claims) in S are true.

(B) The facts in S are relevant to the truth of H.

(C) "All" or "enough" of the true relevant facts are in S.

(D) All told, these speak for the truth of H.

There is currently no dearth of what turn out to be very similar guides that tell you how to evaluate effectiveness claims. Here is a sample:

IARC: International Agency for Research on Cancer

SIGN: Scottish Intercollegiate Guidelines Network

U.S. Department of Education What Works Clearinghouse

USEPA: US Environmental Protection Agency

CEPA: Canadian Environmental Protection Act

Cochrane Collaboration

Oxford Centre for Evidence-Based Medicine

Daubert decision (US Supreme Court).

How do these guides help with A.–D.? First they make a bad presupposition about relevance (B): The only kind of fact seriously relevant to predicting that a policy will work for you is that the policy worked in some studied situation. Then they focus on A., the "quality" of proffered facts of this kind – how likely are they to be true? And they take an odd view about that. Only comparison studies are allowed to count as good evidence that the policy worked in a studied situation, with RCTs as the best among comparison studies. Econometric models, for instance, which can under specifiable circumstances produce good support for causal claims, are not even considered; nor is derivation from well-established theory (Cartwright 2007; Reiss 2005; Fennell 2007a, 2007b). Then the advice given explicitly about B is generally of little practical use. They most often ignore C. And they are very weak on D.

Here for instance is the quality grading scale from SIGN, which is used by NICE (the National Institute for Clinical Excellence) to set best health practice in the UK:

LEVELS OF EVIDENCE

1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
1 -	Meta-analyses, systematic reviews, or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies
	High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2 -	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Non-analytic studies, eg case reports, case series
4	Expert opinion

Notice that, as I remarked, what are ranked are study designs for establishing claims about the efficacy of a policy or treatment in a studied population, and they are all, bar the strange last entry, comparative studies.

The marks from 1++ to 4 are then to be used to help grade policy predictions, presumably to grade them according to how likely they are to be true (or perhaps to be "likely, failing good reason to the contrary"). The grades are assigned thus (Scottish Intercollegiate Guidelines Network 2008):

GRADES OF RECOMMENDATION

A	At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population; or
	A body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
B	A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results; or Extrapolated evidence from studies rated as 1++ or 1+
C	A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or
	Extrapolated evidence from studies rated as 2++
D	Evidence level 3 or 4; or
	Extrapolated evidence from studies rated as 2+

So a really good RCT testing the policy on a study population can be enough to warrant confidence that the policy will work for you – at least it can, on this grading scale, if that RCT is "directly applicable" to your population. That is, it is *relevant* to your population. But when is a well-established claim that the policy worked somewhere – in some study population – relevant to your prediction that it will work for you? The U.S. Department of Education is equally vague. They tell you that strong evidence for your policy is two or more high quality RCTs in "settings similar to that of your schools/ classrooms". The only elaboration later adds four lines – trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students.

These two grading schemes address C. and D. together in one fell swoop. The advice is, "Take the best". If you have two good RCTS, that's enough to go with the policy; you don't need to look at the rest of the evidence. It looks then as if this is advice to reject C and do D an easy way. Other guides take C. more seriously and suggest a judicious consideration of "lower grade" evidence as well, allowing that sometimes lower grade evidence that the policy failed in some study populations should dilute your confidence that the policy will work for you (eg, The Grading of Recommendations Assessment n.d.). But how do you "weigh it all up together"? And isn't weighing already perhaps a misleading way to put the problem? On this issue I think no-one offers very good, practicable advice; and I'm afraid I can't either.[4]

What justifies these answers? There is little grounded discussion. Most of what there is focuses on the grading schemes, in defense of putting RCTS way out ahead. Only these, we are told, control for unknown confounders. But how did unknown confounders enter? This is clearly a remark somewhere very deep in the middle of an argument. I propose to start back at the beginning of the argument, to offer a theory of what counts as evidence from which we can construct answers.
The simple theory I propose does not help much with C. and D. But it has immediate consequences for B., *relevance*, and this is where I shall concentrate. Already, just looking at relevance we will get a very different account of the role of RCTs in warranting effectiveness predictions than we see in these guides and that is suggested by the remarks of Duflo and the GLA document cited in my section **1**.

4. A simple theory of relevance for predicting effectiveness
This theory has two straightforward claims. The first theory claim is

> **TEE 1.** X as implemented will contribute positively to the production of Y in situation S iff
>
> **TEE 1.1.** There is a causal principle that holds in S from implementation till time of outcome in which X figures as a cause of Y,
>
> and
>
> **TEE 1.2.** All the factors that are required in that principle for X to contribute to Y obtain in S at the required times.

TEE 1 is based on two presuppositions:

Law-governedness: If X is to contribute to the production of Y it must do so under the governance of a causal principle.
Analyticity: Causal principles generally allow different kinds contributions to the same effect from different distinguishable sources, some positive, some negative.

The first presupposition supports theory claim TEE1.1. Causings don't happen by accident, by mere hap, but in accord with causal principles at work in the situation. These need not be deterministic. They could for instance be probabilistic causal principles or expressions of causal powers, which may not have probabilities attached. For simplicity though I will use deterministic principles here to keep formal complications to a minimum.

I do not want to endorse the claims that causings happen under principles as a universal truth and, personally, I do not think it is true. But I take it that I am in the minority. More importantly, it is happening under a principle that ensures that a causing can be relied on to happen. It is this that makes effects predictable; and for policy, it is predictability that matters.

Analyticity supports theory claim TEE 1.2. As with law-governedness, I do not take analyticity to be a universal truth and again I personally have argued that sometimes a kind of causal holism that counters it is true. But analyticity is at the base of much scientific method and it is the standard assumption in physics and economics and in large swathes of biology. It may not be so reasonable to assume in social contexts but in that case again policy prediction will be exceedingly difficult, as indeed the Historical School scholars who favored holism argued. So as with law-governedness, I shall take analyticity to be true for the kinds of policy causes that we can sensibly make predictions about. In that case we can suppose that the kind of causal principles that will produce predictable policy outcomes can be represented like this:

$$Y\ c= C_1 + \ldots + C_n - P_1 \ldots - P_m.$$

where the C's and P's are complex combinations of empirical factors and the meaning of "+" can vary. (For instance, when it comes to forces "+" will represent vector not scalar addition.[5])

Principle L from section **2** is just a shortened form of this that focuses on terms that show X as a cause of Y with all the remaining terms gathered together in W.

To complete my simple theory we need only add the second theory claim:

> **TEE 2**: The facts relevant for predicting "X as implemented will contribute positively to the production of Y in S" are those that must obtain in S if that claim is to be true.

So we are warranted in predicting that X will make a positive contribution in S to the extent that we are warranted in assuming:

> **Rel 1.** There is a causal principle that holds in S from implementation till time of outcome in which X figures as a cause of Y,
>
> and

[4] For a survey of problems on evidence amalgamation see Stegenga 2009 and forthcoming .

[5] For examples of other methods of combination see Cartwright 1999, 54.

Rel 2. All the factors that are required in that principle for X to contribute to Y obtain in S and at the right times.

These are the two kinds of facts that are *directly relevant* to predicting a positive contribution from X. Other facts can be relevant – indirectly – by speaking for the truth of either of these.[6]

5. RCTs and Relevance 2

Let us start with Rel 2. since that is easiest and my remarks here are probably familiar. As earlier I shall suppose that laws of form L govern the production of Y for individuals u in the situations of interest, represented by *S*:

L: $Y(u)\ c= \alpha(u) + \beta(u)X(u) + W(u)$.

Causes are INUS conditions. From that it follows that two kinds of facts are directly evidentially relevant to predicting "X will contribute positively to the production of Y in S":

Rel 2.1. Facts about which *are* the factors that regulate whether X contributes to Y in S, and by how much, under a law that governs the production of Y in S,

and

Rel 2.2. Facts about *which of those will be present* in S and at the right times were the policy to be implemented as envisaged.

When laws are expressed in the familiar form of L, facts that are mentioned in Rel 2.1. are represented, in one fell swoop, in β. It is important not to be misled by β's simple form into taking it as a constant or as a simple random variable. β will generally represent a complex function of further factors that together fix whether and how much X contributes to Y. Also keep in mind that there might well be different sets of complexly interacting factors, any one of which allows X to contribute to Y. So β looks like this: $\beta = f_1(z_{11},\ldots,z_{1n}) + \ldots + f_m(z_{m1},\ldots,z_{mp})$.

The quantities represented by *z*'s are sometimes called "confounding factors". That term though is often applied to the quantities represented in W as well. I suggest instead calling the z's that must act in tandem with X to produce a contribution to Y, the *support factors* for X with respect to outcome Y. To make good predictions about the effects of X on Y in S you need first to identify what these are in laws that obtain in S and then to ascertain which of these will in fact obtain in S were the policy to be implemented. The usual advice you receive does not mention the support factors explicitly. But there is an assumption about β built into it. Look back to section **3** and the advice there from SIGN or the US Department of Education. You can expect the same contribution in your situation as in an RCT if your situation is "similar to" that in the RCT or the RCT is "directly applicable" to your situation.

So, similar in what respects? Or when is a result "directly applicable? That is not hard to answer, though the guides don't write it out for you. Looking at formula 2.3. for the treatment effect it is clear that you will get the same treatment effect in S as in an RCT only if your situation and that in the RCT share a law in which the same support factors figure for X with respect to Y and the two situations have the same mean value for these. Otherwise it is an accident. And having the same mean is pretty much an accident unless the two share the same distribution of the different values of β, which represent different combinations of values for support factors.

How are you to make any kind of reasoned judgment about whether the two have the same mean or same distribution of combinations of support factors if you have no idea what these factors are and what range of values they will might take in your situation and with what probability? If you adopt the policy you will be betting that your situation has enough of the right combinations of support factors to ensure a positive contribution overall. But you are seldom urged to consider these and standard guides give little or no advice about how to evaluate the evidential weight of claims offered in support of proposals about them. Edward Leamer, famous for his paper "Let's Take the Con out of Econometrics" (Leamer 1983), in discussing RCTs expresses the same worry about this that I have: "If little thought has gone into identifying these possible confounders, it seems probable that little thought will be given to the limited applicability of the results in other settings"(Leamer 2010) .

What if some thought has gone into it and you have reason to believe you have identified a number of the necessary support factors for X to contribute to Y – the Z's? Then the advice in the guides is poor advice. For you do not really want your situation to reproduce the same distribution across β as in the RCT. Rather you want a distribution that is heavy on the combinations of values of Z's that boosts the input of the value of X you propose to implement and is light on the values in which that X value contributes little or nothing, or worse, is harmful.

Where does this leave us with respect to the relevance of RCTs to our policy predictions? Look at Rel 2.1. RCT results will be indirectly evidentially relevant if they help support claims about what the requisite supporting factors are in our situation. And this they can do. Suppose we hypothesize that Z is a support factor and that $Z = z_g$ is a good value for us – ie, it boosts the contribution of the X value (say *x*) we propose to implement, and we hypothesize that $Z = z_b$ is a bad value. If we can find an RCT situation that we have good reason to believe shares laws for the role of X in the production of Y with our own then we can use an RCT to test this hypothesis, by setting the treatment as $X = x$ and $Z = z_g$ and the control as $X = x$ and $Z = z_b$. It is important to note however that this

[6] We may not be interested just in whether X would contribute positively but in how much contribution we would get from inputting a given value of X. In this case we are not just concerned with whether positive β values obtain in our population (or whether enough positive ones obtain to outweigh the effects of those producing negative contributions) but also with what the mix of values of β is for us.

test is only relevant for us if we have reason to believe that this RCT situation has the same laws for the production of Y from X as does ours – and that needs warranting.

As to Rel 2.2., RCT results are clearly no help in telling whether various factors that have been identified as support factors for us will obtain in our situation. For that we need local information. Sometimes we can tell by looking; sometimes it can require careful measurement; often – as I will discuss in the next section – it will require careful interpretation of what exactly those factors amount to on-the-ground, in-the-concrete in our situation.

6. RCTs and Relevance 1

If we assume law-governedness, as I did in my second treatment of RCTs in section **2**, then a successful well-conducted RCT provides good evidence in favor of the claim that a causal principle like L in which X figures as a cause of Y held in the study situation. This provides no evidence that X will produce a positive difference in the target unless the target and the study share L. L must be, at least to that extent, a general causal principle. I take it this is what is supposed to be expressed in the claim "X works". But the stretch of L is certainly not addressed in the RCT and for the most part generality cannot be taken for granted. That's because the kinds of causal principles that govern policy effectiveness are both **local** and **fragile**.

These principles are local because they depend on the mechanism or the social organization, what I have called the "socioeconomic machine", that gives rise to them. I have developed this claim for locality at length in many places (Cartwright 1999, 2007). Here let me give one vivid example, far removed from our topic of social science and evidence, the case of a Rube Goldberg machine, say this one:

We can fly the kites repeatedly and in different conditions to determine the exact form of the equation by which flying kites lift the small door. The equation may look like this:

Size of door opening $c = \beta$(height of kite) + W.

But we cannot take this principle very far. Kites do not very generally open doors. This causal principle is true but local. This, as I said, is an old theme of mine. I have returned to in recent work because of its central importance for successful prediction about policy interventions.

Some economists are very clear about locality. The Chicago School notoriously used it as an argument against government intervention: The causal principles that governments have to hand to predict the effects of their interventions are not universal. They arise from an underlying arrangement of individual preferences, habits and technology and are tied to these arrangements. Worse, according to the Chicago School, these principles are *fragile*. When governments try to manipulate the causes in them to bring about desired effects, they are likely to alter the underlying arrangements responsible for those principles, so the principles no longer obtain (Lucas 1981).

Or, British econometrician Sir David Hendry (Clements and Hendry 2008; Hendry 2006) urges the use of simple "quick catch-up" models for forecasting rather than more realistic causal models because the world Hendry lives in is so fluid that yesterday's accurate causal model will almost certainly not be true today. JS Mill had similar views (Mill 1836/1967, 1843/1973). Economics cannot be an inductive science, he argued, because underlying arrangements are too shaky; there's little reason to expect that a principle that has held over some period or in some place will hold at a different period or in a different place.

For purposes of policy we want to predict the truth of a singular counterfactual – "This program would work for us, given where and how we would implement it". Nature fixes the outcomes of our policy interventions by working through her own casual principles for the situation. We can try to follow her lead but we will need causal principles *appropriate to our situation*. Our situation will almost certainly have a different complex of causal factors present than any study populations. And implementation generally produces even more differences. We seldom manage to introduce the intended intervention by itself; we usually end up changing lots of other causally relevant factors as well. (Recall the brief discussion of the importance of factors in W in section **2**.) Even more fundamentally, our situation may well have a different underlying structure than the study situations and thus be subject to *different causal principles*; and even if the structure wasn't different to begin with, our actions can alter it, as the Chicago school warns.

Because causal principles are local you can't just take a causal principle that applies here, no matter how sure you are of it, and suppose it will apply there. Perhaps you think – as many other economists and medical RCT advocates seem to – that the different populations you study, here and there, are

more likely to share causal structure than not. That's fine. But to be licensed in that assumption you need good evidence and generally varied in form: evidence about the nature and stability of the structures involved and evidence about the nature and stability of the causal principles they give rise to.

This is a quarrel I have with Judea Pearl (Pearl 2000), who has done such marvelous work on causal inference. I worry about the comprehensiveness of his methods, not their validity. Pearl offers a complete methodology from hunting to using causes. First he provides a general way to represent causal principles; I believe he maintains that his representations are general enough to treat any kinds of causal principles we are familiar with. I don't quarrel with this here. Second he offers a detailed semantics for inferring singular counterfactuals from casual models of this form. Nor do I quarrel with this. Third he points to reliable methods like causal Bayes-nets and RCTs for inferring causal principles from probabilities. Though we probably disagree about how widely the assumptions hold that are necessary for these methods to be valid, I agree that the methods are both powerful and often reliable. The scheme is ideal. We have trustworthy methods for going from data to model and from model to prediction. So the predictions are well supported by the empirical evidence.

The problem is in the joining-up. We need reasons to suppose that the causal principles that produced the data in the studied situation are the same as those that will produce the outcomes we want to predict in the target situation. But we seldom have such guarantees. The probabilistic methods that Pearl and others endorse for discovering causes can provide good descriptive accounts of the network of causal relations that obtain in various populations. These can be a part of the evidence base for the more basic science that allows us to predict what the causal principles might be in new situations.

But simple induction, even if the models are for what is supposed to be the "same" population, is seldom a good tool of inference – and to be warranted in using it, we need good reasons to believe we are studying an entrenched structure. Otherwise for new situations we need to predict new principles and we can't do this by collecting statistics on populations in the not-yet-existing situations. We can though sometimes do so with an understanding of underlying mechanisms and how they interact to generate new causal principles. But for this we need theory, not necessarily grand sweeping theory but theory none-the-less, and in consequence we need the large and tangled confluence of evidence and hypotheses that go into building up and supporting reasonably reliable theory. Of course theory is hard – and unreliable. Simple induction is far easier. But it requires stability and stability is hard to come by. Without at least enough theory to understand the conditions for stability, induction is entirely hit or miss.

Here are the remarks of a pair of other economists who like me stress the importance of finding shared laws if study results on a policy or program are to serve as evidence that the policy will be effective in a target setting:

> Structural analysis gives us a way to relate observations of responses to changes in the past to predict the responses to *different* changes in the future.
>
> It does so in two basic steps: First, it matches observed past behavior with a theoretical model to recover fundamental parameters such as preferences and technology. Then, the theoretical model is used to predict the responses to possible environmental changes, including those that have never happened before, under the assumption that the parameters are unchanged (Nevo and Michael 2010).

To find shared laws I do not think we always have to go all the back to first principles, as the talk of "fundamental parameters" and "theoretical models" suggests. But if we are to use results from a study situation as evidence for predictions about a target, we had better have reason to believe the study results depend on a law that is at least wide enough to cover both.

One way to find shared principles can be to "climb up the ladder of abstraction", to provide more abstract descriptions of the cause and effect factors at work in the study. This is a good idea because laws that hold relatively widely generally involve abstract features. So there is a very rough correlation. The higher the level of abstraction of the features in a principle the wider is its range of applicability. For example, the sun causes the earth to move around it in an the orbit we observe yearly; a large mass causes smaller ones to move around it in elliptical orbits; a mass of size M causes objects at separation **r** from it to accelerate at $GM/\mathbf{r}^2$. Each of these principles is true, and each is true of the earth, which is the original object of focus. But each involves more abstract features than the one before. And each has a wider domain of application than the one before.

These layers of principles can all be true at once, and all apply at once to the earth, because of a simple fact about the abstract and the concrete. Abstract features are always instantiated in more concrete ones. Fables and morals often relate in just this way. Consider a favorite of mine by the German Enlightenment thinker and playwright, GE Lessing (Lessing 1759/1967):

> A marten eats the grouse.
>
> A fox throttles the marten; the tooth of the wolf, the fox.
>
> Moral: The weaker are always prey to the stronger.

The moral of the fable teaches a lesson in the abstract, the fable shows what it amounts to in one or more concrete cases.

The same is very often the case in the social sciences. Economic agents do not always act so as to maximize their income, nor their leisure, nor their consumption, nor the educational levels of their children, nor anything else in the concrete. But perhaps for the most part economic agents act so as to maximize their

expected utilities and when we see them acting for income or leisure this is what, on that occasion, constitutes their utility.

It is just this assumption that underwrites a good many social science experiments. Consider economics. We engineer situations to ensure as much as possible that for the nonce at least, the only source of utility is, say, money to be won in the experimental game. Then we look to see whether, if the monetary rewards are structured like those in a prisoners' dilemma game, agents play the antisocial equilibrium solution predicted from the principle that agents act to maximize their expected utility. If they do, the experimental results are not only an instance of the principle "In a prisoners" dilemma game, agents both defect' but also "Agents act to maximize their expected utility." Conversely, if agents in the experiment cooperate it is not just a challenge to the principle about what should happen in a prisoners' dilemma game, but a challenge to the fundamental principle of utility maximization as well.

To get a shared lesson from a study like an RCT it is often best then to couch that lesson in far more abstract terms than those in which it is carried out. But there is a problem. How do you know when you have made the right associations between the abstract and the concrete?

Sometimes the identification can be easy. Here's a case I discuss in more detail elsewhere (Cartwright 2011). There was good evidence that a nutritional counseling program for mothers in the Indian state Tamil Nadu improved the nutrition of their young children. Yet a similar program did not succeed in Bangladesh. The principle "nutritional counseling in mothers improves their young children's nutrition" was too local to cover Bangladesh as well. The after-the-fact evaluation of the Bangladesh program indicated that a good part of the reason the program didn't work there was that often mothers neither did the shopping – the men did it – nor controlled food distribution in the family – their mother-in-law did that. Let's take this account for granted for purposes of illustration.

The information from the evaluation plus the background knowledge that prompted the nutrition program in the first place[7] make it a good hypothesis that there is a more abstract principle that holds in both Tamil Nadu and Bangladesh: Nutritional counseling for those who procure the family food, control its distribution in the family and reflect concern for a child's nutritional welfare in doing these improves the child's nutrition. In India mothers satisfied the complex description in this (hypothetically) shared principle, though not in Bangladesh. And in this case it would not be hard to verify that in Bangladesh "mother" does not do so – just go to the market and notice that all the food shoppers are male.

Other cases will be more problematic. The prisoners' dilemma experiment is designed to make easy the identification of money won in the game with utility but what tells us what is to count as utility in most naturally occurring situations? I take it that sometimes we will be in a position to make and defend identifications and sometimes not; it will depend on our background knowledge, both local and general. It is important to stress that this is not the kind of knowledge that RCTs are good at securing. It requires a different kind of scientific backup and exactly what kind – or better, "kinds" – is required can differ from case to case.

This points out a great shortfall in helpful advice. In general evidence hierarchies like those mentioned in section **3** only rank methods for producing evidence that a program works somewhere, with no advice at all about how to judge when evidence proffered for identifying features across levels of abstraction is likely to be sound and strong.

Let us return finally explicitly to the lessons about where RCTs are relevant. RCT results are relevant only to situations where the effect is produced under a principle shared with the study situation. The very methodology of the RCT tends to restrict its range of relevance. In order to ensure that all members of the treatment group receive the same treatment it is important for a proper RCT that the program under test – the treatment protocol – be very precisely specified. This means describing the treatment in very concrete terms and the very concrete features picked out by the protocols are likely to figure only in very local principles. Yet the very same RCT results can at the same time be evidence for a principle connecting more abstract features that has far wider coverage – if only we can import the web of background knowledge that it takes to recognize those features and support the breadth of the more abstract principle. The point is that sometimes we have such knowledge, or at least a body of evidence that provides reasonable support for it.

Many RCT advocates, however, urge trusting to nothing but what can be established by an RCT. The irony is that this advice undermines the usefulness of RCT results and it is at any rate impossible to follow. To present results some choice must always be made about what the relevant treatment and effect features are. That is a choice that cannot be underwritten by RCTs. Rather it depends on the complicated kind of consilience of theory and empirical studies that is always necessary to grasp what features, among the panoply on offer, are linked in the kinds of regular ways we describe in our causal principles.

Perhaps we get carried away by drug trials, where we suppose a vast amount of biochemistry and knowledge of human physiology picks out what are the widely applicable relevant treatment features, background knowledge that we take to be so secure that we can ignore its role in warranting our choice of features to figure in the causal principles that we take the RCT to test. Whether this is true about drugs and certain other medical treatments, it is surely very doubtful about social interventions. It is almost a truism in social science that one

[7] This might include, for instance, well-evidenced claims that in both places mothers believed in eating down during pregnancy or conventional child rearing habits forbade certain nourishing food to children, like fish.

and the same thing, concretely described, can have very different meanings in different social, cultural and economic settings, and hence have very different effects.[8]

What then of advice that urges that exactly the same protocol as in the RCT should be satisfied if one is to expect the same kinds of results in a new setting as in the RCT? For instance the Greater London Authority *Standards of Evidence* document claims: "… it is now established that programs that are delivered with what is called 'fidelity' – meaning they are implemented as intended by the program designers – achieve the best results"(Greater London Authority n.d.).

On the one hand, this can be good advice. Policy makers quite reasonably often try to cherry pick through the features of a program to implement only ones that are cheaper, or more feasible or more politically/culturally acceptable in hopes that they will still get reasonably close results to those in the RCTs. This can be a very bad idea for often program features are interactive and program designers have been at pains to include enough of the right combination to ensure that the program, taken as a whole, will generate positive contributions, but if factors are omitted, a positive contribution is unlikely – that is, they have worked to ensure that many of the main members of what I have called the requisite "support team" is built into the program design.

An example might be the Nurse-Family Partnership (Olds 2006; Olds et al. 2003) that has been used in a number of US cities and is now being introduced into the UK to improve pregnancy outcomes, child health and development, and parents' economic self-sufficiency. It involves a heavy program of prenatal and infant home visiting and is thus costly – though, as designers claim, not relative to the costs saved from many of the later problems averted by the program, let alone the suffering averted. The designers will not sell the license for this program unless it is to be taken up in its entirety – though they are concerned to pursue ways to adapt the program to make it more suitable to, for instance, use in Birmingham, which is where it is first being tried in the UK.

So it can be a good idea to stick to the RCT protocol. On the other hand it can be a very bad idea for the reasons I've rehearsed: It's not the protocol that matters more widely but rather something more abstract that the protocol instantiates in the RCT situation.

How do you judge when it is good advice and when bad? As I said, that takes a good network of theoretical and empirical knowledge. A policy that worked somewhere will not work for you unless there's a shared principle that governs the results in both the study situation and your own. If you don't have good reason to believe there is one, and that you have correctly identified what the appropriate concrete form of the treatment

[8] For another example involving child welfare practices see Cartwright 2012.

is under that principle in your situation, then you are betting when you set out on a policy course and probably betting at unknown odds. Sometimes you need to do this. But it's best to acknowledge it and manage the uncertainly as best possible, not act as if you have warrant that you lack.

7. Summary

There are two kinds of facts directly relevant to predicting "X will contribute to Y in S": the facts that

> Rel 1. The production of Y in S is governed by a law L in which X appears as a cause of Y.
>
> Rel 2. All the factors necessary under L to support X in producing a contribution to Y obtain in S.

Other facts are relevant – indirectly – if they are relevant to establishing either of these facts. This includes the kind of fact that an RCT can lend support to – that X caused Y in a study situation, or that the effect difference for X with respect to Y was positive in a study situation.

What then in more detail of the evidential relevance of RCT results to effectiveness predictions?

A positive effect size for treatment X and outcome Y in an RCT in situation R is directly evidentially relevant to "R is governed by a law, say L, in which X causes Y." The positive effect size *can* be indirectly relevant to facts about the laws in another situations S – "S is governed by a law in which X causes Y" – but its relevance is conditional on the fact that R and S share L. That is, a positive effect size in R is relevant to whether X figures as a cause of Y in S if, but only if, R and S share L. For evidence-based policy we should have good reason before we assume this. And RCTs will be hard pressed to provide that reason. Even if the RCT were conducted on a sample of the target population, samples can be misleading. Equally important, one must suppose that the causal structure does not change from the time of study till the time the policy begins to work its effects. That is an empirical hypothesis that may or may not be true and that should not simply be assumed without reflection and without reason to back it up.

Not only do RCTs not tell us that two situations share a causal structure. They also fail to tell us what the operative factors in the causal principles are. S may share with the study situation R a general causal principle under which the results in the study are produced, but the causes in the shared principle need not be the ones described in the protocol of the study. The protocol may pick out causes at too low a level of abstraction for sharing and the very same protocol carried out in S may not constitute the same cause that it constitutes in Y. Again, the RCT can be indirectly relevant, but only conditionally, in this case conditional on the fact that the protocol in the study and the proposed policy both instantiate the causes in a shared causal principle.

As to Relevance 2, RCTs are not relevant to "All the supporting factors required by L for T to contribute to O obtain in S." That requires

different kinds of evidence, more local and different in kind. It requires studies aimed at establishing what features are there in S, not ones geared to establishing causal connections. RCTs can be relevant to identifying what the supporting factors are. But again, only conditionally: If – but only –if the study and target situations share the relevant causal structure. And again, for evidence-based policy this should not be assumed without reflection and without reason.

Overall, the lesson is simple. It is a long road from an RCT that evidences the fact that a policy works somewhere to the prediction that the policy will work for us. A lot of different kinds of facts requiring evidence of different kinds to support them must be in place before the road is secure, or secure enough for us to bet on it given the costs and benefits of success and failure. That makes policy predictions dicey – but then that is something to expect. Many of the facts we need to establish are sometimes within our grasp or can become so with reasonable effort. Sometimes they aren't and we need to hedge our bets as best we can. But in any case, the more of the road we support, the more likely it is that our inferences will go through.

References

Bohrnstedt, G. W. , and B.M. Stecher, eds. 2002. *What we have learned about class size reduction in California*. Sacramento, CA: California Department of Education.

Buckley, Cara. 2010. To test housing program, some are denied aid. *The New York Times*, December 8. http://www.nytimes.com/2010/12/09/nyregion/09placebo.html?scp=1&sq=To%20Test%20Housing%20Program,%20Some%20Are%20Denied%20Aid&st=cse (accessed December 26, 2010).

Cartwright, Nancy.1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

——— 2007. *Hunting causes and using them*. Cambridge: Cambridge University Press.

——— 2009. Evidence-based policy: What's to be done about relevance, models, methods, and evidenc. Proceedings of the 38th Oberlin Colloquium in Philosophy, edited by M. T.-J. (ed.): *Philosophical Studies* 144 (1): 127-136.

——— 2011. Evidence, external validity and explanatory relevance. In *The philosophy of science matters: The philosophy of Peter Achinstein*, edited by G. J. Morgan. Oxford: Oxford University Press.

——— 2012. Will this policy work for you? Predicting effectiveness better: How Philoshophy helps. *Philosophy of Science* 79.5 (2012): 973-989

Cartwright, Nancy, and Jacob Stegenga. 2009. Towards a theory of evidence for effectiveness for evidence-based policy. In *NRC Conference on evidence-based policy: International experiences*. Paris. http://www.oecd.org/dataoecd/36/62/42382313.pdf (accessed December 30, 2010).

Cartwright, Nancy, and Jacob Stegenga. 2011. A theory of evidence for evidence based policy. In *Evidence, inference and enquiry*, edited by Philip Dawid, William Twining and Mimi Vasilaki. London: British Academy Publication.

Clements, Michael P., and David .F. Hendry. 2008. Economic forecasting in a changing world. *Capitalism and Society* 3 (2):1-18.

Fennell, Damien. 2007a. Why and when should we trust our methods of causal inference? Lessons from James Heckman on RCTs and structural models.

Contingency and dissent in science technical report 06/07. London: Centre for Philosophy of Natural and Social Science, LSE.

——— 2007b. Why functional form matters in structural models in econometrics. *Philosophy of Science* 74 (5):1033-1045.

Greater London Authority. n.d. www.london.gov.uk

Heckman, James J . 2001. Econometrics, counterfactuals and causal models. Keynote Address at International Statistical Institute, at Seoul, South Korea.

——— 2005. The scientific model of causality. *Sociological Methodology* 35:1-97.

Hendry, David F. 2006. Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics* 135 (1-2):399-426.

Holland, Paul W., and Donald .B. Rubin. 1988. Causal inference in retrospective studies. *Evaluation Review* 12 (3):203-231.

Leamer, Edward E. 1983. Let's take the con out of econometrics. *American Economics Review* 73 (1):31-43.

——— 2010. Tantalus on the road to Asymptopia. *Journal of Economic Perspectives* 24 (2):31-46.

Lessing, Gotthold Ephraim. 1759/1967. *Abhandlungen uber die Fable*. Stuttgart: Philipp Reclam.

Lucas, Robert E. 1981. Economic policy evaluation: a critique. In *Studies in business cycle theory*, edited by R. Lucas. Oxford: Basil Blackwell

Mill, John Stuart. 1836/1967. On the definition of political economy and on the method of philosophical investigation in that Science. In *Collected works of John Stuart Mill*. Repr.Toronto: University of Toronto Press.

——— 1843/1973. On the logic of moral sciences. In *Collected works of John*

Stuart Mill. Repr.Toronto: University of Toronto Press.

Nevo, Aviv, and Whinston Michael. 2010. Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives* 24 (2):69-82.

Olds, David L. 2006. The Nurse-Family Partnership: An evidence-based preventive intervention. *Infant Mental Health Journal* 27 (1):5-25.

Olds, David L., Peggy Hill, Ruth O'Brien, David Racine, and Pat Moritz. 2003. Taking preventive intervention to scale: The Nurse-Family Partnership. *Cognitive and Behavioral Practice* 10 (4):278-290.

Pearl, Judea. 2000. *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Reiss, Julian 2005. Causal instrumental variables and interventions. *Philosophy of Science* 72 (5):964-976.

Scottish Intercollegiate Guidelines Network. 2008. *SIGN 50 A guideline developer's handbook*, *Annex B: Key to evidence statements and grades of recommendations*. www.sign.ac.uk/pdf/sign50.pdf (accessed December 30, 2010).

Stegenga, Jacob. 2009. Robustness, discordance, and relevance. *Philosophy of Science* 76 (5):650-661.

——— forthcoming. Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*.

The Grading of Recommendations Assessment, Development and Evaluation Working Group. n.d. www.gradeworkinggroup.org.

Worrall, John. 2002. What evidence in evidence-based medicine? *Philosophy of Science* 69 (S3):S316-S330.

——— 2007. Why there's no cause to randomize. *The British Journal for the Philosophy of Science* 58 (3):451-488.

SECTION III

EVIDENCE IN THE ABSTRACT:

A GENERAL THEORY OF EVIDENCE WHERE RIGOR MATTERS

3.1 DOES ROUSH SHOW THAT EVIDENCE SHOULD BE PROBABLE?

(With Damien Fennell from Synthese 2010)

1. Introduction

Evidence has always been a central topic in the philosophy of science. But with debates raging throughout the US and the UK about what counts as evidence for evidence-based policy, the topic has become one of immediate practical importance. This makes very timely Sherrilyn Roush's recent efforts to characterize evidence.

Roush presents her account near the end of her detailed development of a tracking theory of knowledge, now one of the central topics in epistemology. Even if tracking is not the correct or the only good account of knowledge, it would certainly be a plus to have a theory of evidence for which it could be assumed that anything that counts as evidence is a source of tracking knowledge. We shall here look into the question of how her theory of evidence dovetails with her tracking view of knowledge.

Our primary topic, however, is another important claim that Roush defends. Roush claims for her account the special virtue that it "explains why better evidence makes knowledge more probable" (Roush, 2005, 185). This seems a truism: What we mean by better evidence for h is evidence that makes h more probable. But this is not what Roush means. For her, "better evidence" is evidence that has a higher probability. If e is to be evidence for anything, she maintains, it is ideal that P(e) be high.

This conclusion matters, and not just for the philosophical explication of the concept of evidence. Now that evidence-based policy is widely mandated, guides outlining what counts as evidence for policy effectiveness proliferate.[1] Roush's demand for high P(e) is at their heart. The guides offer schemes that rank methods for producing evidence according to the degree of certainty that the method confers on the conclusions it produces: High-quality evidence claims are claims produced by methods that make it likely that those claims are true, claims e for which P(e) is high,[2] No policy recommendation can get a top-grade rating unless it has top-ranked evidence claims in its favour. For these guides it is not only "ideal" that P(e) be high if e is to count as evidence; it is necessary.[3]

There is a simple reason why one might adopt this view. It is almost certainly what motivates

[1] For instance evidence evaluation schemes from the Scottish Intercollegiate Guidelines Network, the International Agency for Cancer Research, or the Maryland rules in criminology.

[2] The guides clearly seem to make the assumption that high probability can be assigned to results from reliable methods. While an important issue, this is tangential to concerns here about whether a claim must have high probability if it is to be counted as evidence.

[3] This raises a question as to how to interpret probabilities. Roush claims that her definition of evidence is compatible with both subjectivist and objectivist readings. In section 3 we explore issues relating to the subjectivist interpretation of probabilities when we consider Roush's argument from a Bayesian standpoint. To interpret Roush in an objectivist way, we avoid standard controversies in the philosophy of probability by assuming that e and h both denote event-types. This, however, is not to say we believe that it is straightforward to find an interpretation of probability that makes sense of Roush's definition of evidence or her arguments for probable evidence. Indeed, the difficulties discussed in section 3.2 suggest otherwise.

the ranking schemes and it is one of the reasons Roush herself gives: If e is to be good evidence for h, e should provide good reason to believe h. Surely we shouldn't believe h on the basis of e unless there is good reason to believe e. So P(e) should be high. We shall here accept this line of defence for high P(e) without discussion and concentrate on the rest of Roush's discussion, for she has far more than this to offer. In particular she develops two original, challenging defences of high P(e), both of which open new perspectives on the age-old topic of evidence. The first is based on an interesting mathematical relationship and a related series of graphs and the second on arguments against modelling surprising evidence as evidence with low probability.

We shall argue that these defences do not carry the conclusion. In good part that is because there is not one conclusion in Roush's discussion but three, all expressed in the same words: High P(e) is ideal if e is to be evidence for h. We claim that there are three conclusions because we think there are three different senses of "evidence" at play in Roush's discussion, senses that are important to distinguish independent of their role in the specific issue of P(e). They are –

(1) Evidence as *the ground for knowledge*.[4] In order for e to be evidence for h, e should be an appropriate basis for knowledge that h. The version of "evidence as the ground for knowledge" we find in Roush supposes both that h be true – "as it must be for anyone to know it" (Roush, 2005, 153) and also that e provides grounds for believing it so.[5]

(2) Evidence as a two-place *relevance relation* ("e is evidence for h") between propositions or possible events, in which the evidence is supposed to be relevant to the truth of the hypothesis, without any presumption about whether either the evidence or the hypothesis is true.

(3) Evidence for a hypothesis h as a *lever* to infer P(h), that is, knowledge about the evidence or its probability can be used to deduce informative, previously unknown constraints on P(h), or better, P(h) itself.

The second is the usual topic of confirmation theories and one could take it that Roush's explication is aimed here since she engages with the conventional literature at various points. It is at any rate an important topic, and again, not one just of philosophical interest. Consider hypothesis testing or policy deliberation. Gathering facts, finding out what is true and what is not, conducting experiments, even just sitting and discussing the facts is costly and time consuming. So one wants a concept of evidence that tells what facts bear on the hypothesis in order to decide which ones to find out about, which experiments to run or which facts to let onto the table for discussion. This is looking at evidence from the perspective of the deliberation process, prior to any views about whether what is admitted as evidence provides sufficient grounds for believing the hypothesis, ie, before considerations about issue (1). This perspective also fits particularly nicely with Roush's own concerns, which we separate out as issue (3), that evidence should provide leverage. She does not want P(h) to be presupposed in our attempts to settle if her two central requirements for evidence are met because that would undermine our ability to leverage from the evidence to the hypothesis.

Roush's discussion of high P(e) does not differentiate these three notions, yet P(e) seems to fare differently in each. For sense 1 it seems natural to suppose evidence should have high probability for the trivial reason that e can hardly be the basis for knowledge that h if e isn't itself true, or highly probable, just as the evidence-ranking schemes suppose. But high P(e) should surely be omitted as a criterion for evidence in sense 2. For sense 3, we shall argue, none of Roush's three criteria are necessary.

We look at Roush's defences of high P(e) in section 3, evaluating them both on their own merit and with an eye to disentangling sense 1 and 2 of "evidence". We look at leveraging in section 4. In section 5 we produce a simple counterexample to show that, strictly speaking, evidence for h as defined by Roush does not imply that e tracks h, contrary to her hopes.

2. Roush's graphical analysis and definition of evidence

Roush constructs her definition of evidence from two desiderata. The first is that evidence should *discriminate* between hypotheses. She takes this to mean that if e is evidence for h then $P(e|h) > P(e|\neg h)$, or, in terms of the likelihood ratio ($LR = P(e|h)/P(e|\neg h)$) that $LR > 1$.[6] Roush takes the discrimination condition to be uncontroversial and focuses greater attention on a second desideratum, the *indication* condition, $P(h|e) > 0.5$. The motivation for this is that evidence should, when true, make the hypothesis more likely than not, thus giving us some reason to believe h (rather than its alternative ¬h).[7] These are both well rehearsed conditions, familiar from debates about how to define what we have prised out under the label "evidence as a relevance relation". For present purposes we shall take her arguments for them as sufficient for evidence in a relevance sense since our focus is on her further requirement that P(e) be high. Roush defines evidence so that both

[4] The idea here is similar to Williamson's "E=K" thesis (2000, 185) that evidence is just what we know. This thesis is motivated, as with Roush, by a desire to use evidence to justify belief in a hypothesis. Williamson's treatment is similar to Roush's in other respects as well. For instance, he requires that $P(h|e) > P(h)$ for e to be evidence which is equivalent to Roush's discrimination condition. However he does not adopt a condition similar to Roush's indication condition $P(h|e)>1/2$. Instead Williamson uses the fact that evidence is knowledge and the requirement that $P(h|e)>P(h)$ to justify belief in the hypothesis from evidence.

[5] As Roush puts it in her discussion of tracking and evidence "… if h is true – as it must be for anyone to know it – and e tracks h then it is unlikely that e is false. And, if e is false, then because the subject's belief in e tracks e, the subject is unlikely to believe e. Since b(h) tracks b(e), the probability of b(h) given b(e) is low too. All of this suggests that if the subject knows h through this trajectory, then because in order to do that she must believe h, e is likely to be true." (Roush, 2005, 153).

[6] She further invokes a number of authors to argue that the likelihood ratio is the best measure of how good evidence is at discriminating.

[7] Roush notes "… we do not have good reason to believe, or even some reason to believe, a hypothesis is true, if we have no assurance that the posterior probability [P(h|e)] is greater than 0.5" (Roush, 2005, 165).

desiderata are met, with particular emphasis on the indication condition.

R: e is some/good evidence for h if and only if "there is a lower bound greater than 1 on [LR] and a lower bound greater than 0 on P(e) such that P(h|e) is greater than 0."/ "greater than some high threshold appropriate to having good reason to believe" (Roush, 2005, 183).

Despite its roundabout formulation in terms of lower bounds, the definition is logically equivalent to the following simpler definition.

(DEF1) e is *some* [alternatively, *good*] *evidence* for h if and only if
DC *(Discrimination Condition)*: LR > 1
IC *(Indication Condition)*: P(h|e) > 0.5 [alternatively, P(h|e) > a, where a is some chosen level greater than 0.5]

But this definition does not sit well with other statements made by Roush:

"An obvious solution … is to adopt as a second condition for e to be evidence the demand that …P(h|e), be high. *However, that is merely a restatement of our desideratum.*" (emphasis added, Roush, 2005, 166).

Also, the lack of reference to a lower bound on P(e) in (DEF1) ignores the importance Roush attributes to evidence being probable:

R-addendum: "high P(e) is not necessary but is ideal" (Roush, 2005, 183).

This suggests that (DEF1) is not an adequate interpretation of Roush's definition of evidence. An alternative, more suitable definition can be discerned by careful consideration of Roush's graphical analysis, where she uses a series of graphs to explain the connection between lower bounds on LR and P(e) and ***IC***. These are based on an identity that she establishes using the probability axioms:

A. P(h|e) = [LR-P(e|h)/P(e)]/[LR-1].

She points out that ***A*** implies facts about how P(h|e) can increase under special circumstances. The special circumstances are that
(1) LR>1
(2) LR is held fixed
(3) P(e|h) is held fixed.
Note that this implies that P(e|¬h) is also fixed.

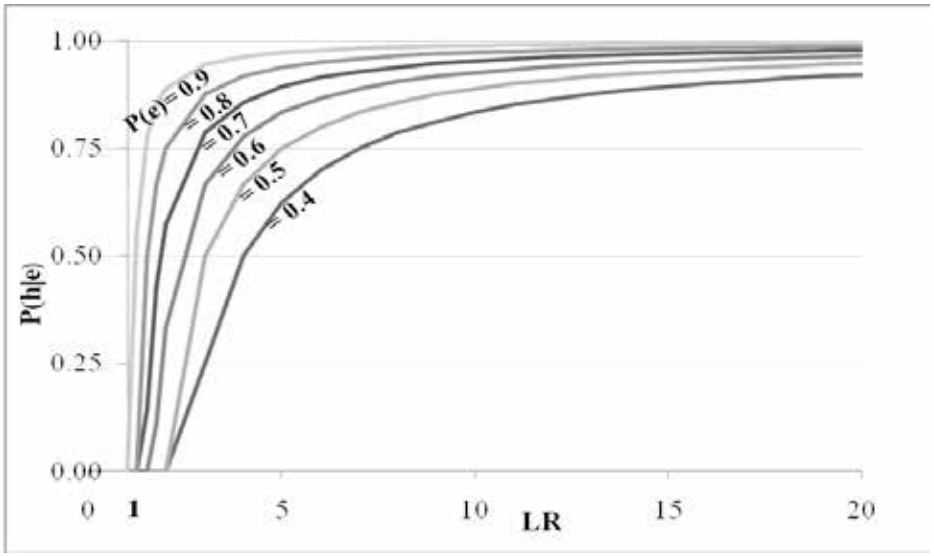
Given these three conditions it follows from ***A*** that

B. P(h|e) increases with increasing P(e).

Roush elaborates on ***B*** defending her advice that P(e) should ideally be high by taking the reader through a series of graphs and formulae.

We summarize her eight graphs in Figure 1.

Figure 1 - P(h|e) vs. LR for various fixed P(e) with P(e|h) = 1



Her graph for P(e) = 0.4 is the one she presents first. On it is displayed a shaded area above the P(e) = 0.4 line in Figure 1, representing the continuum of lines graphing P(h|e) versus LR for the continuum of values of P(e) possible above P(e) = 0.4.

She explains:

"This graph presents a convenient lower limit for the trends that we will see when we increase LR and P(e). The result is this: this surface bounds from below in the [P(h|e)] dimension every graph with LR>1 and P(e)>0.4, and as these terms increase the [P(h|e)] term increases. That is, as LR and P(e) increase above 1 and 0.4 respectively the value of P(h|e), for any given values of P(e|h) and the LR, monotonically rises to 1. We can see why this is by inspecting the equation

P(h|e) = [LR – P(e|h)/P(e)]/(LR-1)

If we suppose that the LR is some fixed value greater than 1, then P(e|h) will be greater than or equal to P(e)… In other words, increasing P(e) with fixed or rising LR will have the effect of increasing P(h|e)." (Roush, 2005, 168)

This means that with fixed LR, high enough P(e) guarantees whatever value of P(h|e) is demanded. Specific values for P(h|e) are necessary on Roush's characterization for e to be some/good evidence for h. Supposing LR> 1, it follows that high enough P(e) is a sufficient condition for a condition (viz. P(h|e) > a) that is necessary for e to be some/good evidence for e. High P(e) is not necessary though, as Roush herself notes in places. We

stress this because we found some comments in her text that could be misleading on this.[8]

Roush's graphical analysis also suggests a corresponding definition of evidence. In her analysis, the lower bounds on LR and P(e) that are sufficient for ***IC*** are independent of the value of P(e|h). Indeed, she states as much: "Through graphing P(e|h), P(e) and the LR, we have found a result that is independent of P(e|h) and depends only on P(e) and LR." (Roush, 2005, 168) Though she does not say this explicitly, this suggests that Roush intends the lower bounds on LR and P(e) in her definition of evidence to be independent of the value of P(e|h). To construct a corresponding formal definition of evidence, it is first instructive to construct lower-bound conditions equivalent to ***DC*** and ***IC***. In the appendix we prove:

C. ***DC*** and ***IC*** if and only if there exist $x > 1$, $y > 0$ and $0<z \leq 1$ such that

(i)LR $\geq x$
(ii)P(e) $> y$
(iii) P(e|h) $\leq z$
(iv) $\frac{x - z/y}{x - 1} = \frac{1}{2}$.

Condition (iii) shows how, given lower bounds on P(e) and LR, whether ***DC*** and ***IC*** are met depends on P(e|h). Roush's apparent desire to construct a definition of evidence in terms of lower bounds that are independent of P(e|h) suggests the following natural move: To stipulate that z=1 so that condition (iii) becomes P(e|h)≤1, which always holds regardless of the value of P(e|h). This suggests defining evidence this way:[9]

(DEF 2) e is *some* [alternatively, *good*] evidence for h if and only if there exist $x > 1$, $y > 0$ such that

(i)LR $\geq x$
(ii)P(e) $> y$
(iii) $\frac{x - 1/y}{x - 1} = \frac{1}{2}$

[alternatively $\frac{x-1/y}{x - 1} = a$ for appropriate $a>\frac{1}{2}$]

(DEF2) fits Roush's roundabout expression and gives P(e|h)-independent lower bounds in line with her graphical analysis. It also makes explicit a trade-off: As the lower bound on LR strengthens, the lower bound on P(e) can weaken and *vice versa*. Given this trade-off, high P(e) is not necessary for e to be evidence, since for any low y there is a sufficiently high x that ensures that e is evidence. Yet for any given x, a higher y raises the lower bound on P(h|e) setting out how more probable evidence is ideal. In these ways the above definition of evidence neatly fits Roush's discussion.

In this definition ***IC*** and ***DC*** are necessary but not sufficient for evidence as can be seen in the following probability distribution:

[8] For instance, when she says "my proposal, then, is that the second condition on evidence, the indication condition, be a lower bound on the value of P(e)". Or just after, "there are three broad questions to ask about this idea…a third is whether high P(e) is plausibly a necessary condition for evidence, since there seem to be counterexamples" (Roush, 2005, 171).

[9] This reading is in accord with suggestions from an anonymous referee.

P(e) = 0.6001, P(h) = 0.8, LR= 2
For this distribution P(h|e) ≈ 8/9. But e is not evidence because the lower bound on P(e) would have to be less than 0.6001. By (iii) this implies that the lower bound on LR would have to be greater than $2\frac{1}{3}$, which is false. This example shows that this version of Roush's definition has an undesirable consequence of ruling out some cases where e is probable and ***IC*** and ***DC*** are met. Why shouldn't e count as evidence in cases like this?

What is missing here is why a definition of evidence should be constructed in terms of lower bounds of LR and P(e) that are independent of the value of P(e|h). Though Roush does not explicitly discuss this, one candidate answer is that it yields a leverage advantage by it allowing one to classify e as evidence without needing to know P(e|h). Roush's discussion of the leverage advantage of using a lower bound on P(e) (Roush, 2005, 170) suggests that this may be her motivation. However, this leverage advantage comes at a cost, since it rules out the case above, which seems paradigmatically evidence on Roush's terms (***IC*** and ***DC*** are met, P(e) is high). Moreover ignorance of P(e|h) also makes it hard to evaluate LR!

Finally, it should be stressed that, though Roush's definition requires the existence of a lower bound on P(e), it does not require that evidence be probable at all, since the lower bound on P(e) can – provided LR is high enough – be arbitrarily close to zero. So, although probable evidence is defended by Roush as ideal, it is not a necessary condition of evidence as she defines it. We think this is an advantage, since we now defend improbable evidence.

3. In defense of improbable evidence

We find in Roush three major arguments that high P(e) is ideal:

*The argument from equation ***A*** and the accompanying graphs.
*An argument against a simple story of Bayesian updating.
*High P(e) has a leveraging advantage for finding P(h).

The first argument seems most suited to a relevance sense of evidence, but we discuss it in section 3.1 more or less on its own grounds without putting weight on our view that there are three different senses of evidence involved in Roush's discussion. The second argument seems geared to knowledge as a grounds for knowledge. Bayesian updating allows that surprising evidence – in the sense of evidence with low probability – can increase the degree of belief in a hypothesis more than non-surprising evidence. Roush's discussion seems to suppose that this is incompatible with her view that evidence should have high probability. We address this in Sect. 3.2. Sect. 3.3 briefly considers other shorter defences Roush offers for high P(e). We take up leveraging in Sect. 4.

3.1 Low probability evidence can satisfy **DC** and **IC** maximally

We assume, for the sake of argument, that high P(e) is a reasonable requirement on evidence as a ground to knowledge. But we

do not see how it can be taken as an ideal way to satisfy the definition Roush offers, which at any rate we think is best seen as a reasonable candidate for defining a relevance notion of evidence. For despite Roush's proposal, a lower probability claim can make for better evidence using her own criteria. Suppose P(e|h) = 1,[10] which is one way to model "h explains e" in the deductive-nomological account of explanation. Then Bayes Theorem reduces to

D. P(h|e) = P(h)/P(e).

Since P(e|h) = 1, it is follows that

E. P(e) = P(h) + P(e|¬h)P(¬h)

and so

F. P(h|e) = P(h) / [P(h) + P(e|¬h)P(¬h)].

Given P(e|h) = 1 it also follows that

G. LR = 1/P(e|¬h).

So lowering P(e) by lowering P(e|¬h) simultaneously produces improvements in LR and in P(h|e), making e better evidence for h on both Roush's criteria. While it is true, as she concludes, that "increasing P(e) with fixed or rising LR will have the effect of increasing P(h|e)" (Roush, 2005, 168), it is equally true that decreasing P(e) with rising LR can have the effect of increasing P(h|e). Thus the graphs hardly provide a strong argument for increasing P(e) in order to satisfy the criteria for evidence.

Not only can lowering P(e) raise both LR and P(h|e), but both conditions ***DC*** and ***IC*** can be maximally satisfied while P(e) takes any value whatsoever. Suppose e is a perfect sign of h; ie, e ≡ h. Then P(h|e)=1 and LR is infinitely high, but P(e) can be as small or as large as one would like. This example has another nice aspect. Whenever there are two independent criteria for the same thing, trade-offs may be required, but here a trade-off is avoided. In this case (or any case with fixed P(e|h)) less probable evidence can be better evidence by both criteria at once.

We should also note that Roush's graphical arguments for high P(e) depend on the asymmetry with which she treats the two independent criteria for evidence. Suppose e is "candidate" evidence for h in the sense that ***DC*** is well satisfied (ie, LR is high). Then high P(e) is sufficient for the satisfaction of ***IC***. But the exactly symmetric claim is not true. Suppose e is "candidate" evidence for h in the sense that ***IC*** is well satisfied. Then it is not true that high P(e) is sufficient for the satisfaction of ***DC***.[11] So high P(e) is useful for obtaining high P(h|e) when LR is sufficiently high, but high P(e) is not sufficient for high LR when P(h|e) is high. Yet there seems to be no special reason for considering either criterion differently from the other.[12]

3.2 A Bayesian defense of improbable evidence and a frequency defence of probable evidence

As a prelude to our arguments, we first set out two distinct analyses on how the probability of evidence relates to the probability of the hypothesis. The first is Roush's while the second is an analysis often presented to support the conventional claim that evidence with a lower probability makes a hypothesis more probable (all else being equal) than evidence with a higher probability.

Section 2 described Roush's argument that if LR is sufficiently high a lower bound on P(e) is sufficient for a lower bound on P(h|e). Importantly, given Roush's constraints – P(e|h) fixed and LR fixed (>1)) – P(e) increases if and only if P(h) increases.

A second analysis, one conventionally used in discussions of the greater confirmation power of surprising evidence, [13]follows from Bayes Theorem:

P(h|e) = P(h)P(e|h)/P(e)

In this case, assuming P(e|h) and P(h) are fixed, P(h|e) increases as P(e) decreases. In this case, LR must increase and P(e|¬h) decrease when P(e) decreases. Labelling evidence that has lower probability as "more surprising", this result shows that the more surprising e is, the higher P(h|e) is.

These two analyses can be summarised as follows:

Since both analyses follow from the

Table 1 – Two analyses of relationship between P(e) and P(h|e)

Analysis	*Fixed Factors*	*Change to P(e)*	*Resulting change in P(h|e*	*Resulting changes in other factors*
Roush	LR (>1), P(e|h)	P(e) increases	P(h|e) increases	P(h) increases, P(e|¬h) fixed
Conventional	P(e|h), P(h)	P(e) decreases	P(h|e) increases	LR increases, P(e|¬h) decreases

[10] Similar examples can be generated for any fixed non-zero value of P(e|h). Note that given P(e|h) = 1, e is evidence if and only if it satisfies DC and IC.

[11] See theorem 2 in the appendix.

[12] We have not here rehearsed Roush's arguments for DC and IC but we don't find anything in them that gives a reason to treat the two in this different way.

[13] See, for example, Howson and Urbach (2005, p.97).

probability axioms, there is no contradiction between them despite their apparently conflicting conclusions as to the relationship between changes in P(e) and changes in P(h|e). As the table makes clear, the difference is due to different factors being held fixed.

This is just arithmetic with probabilities. However, both analyses are used to make arguments as to the significance of the probability of evidence. The conventional analysis is used to argue, via Bayesian updating, that the more surprising evidence is, the more confirmation it lends to the hypothesis once learned. This clearly involves evidence in the sense of grounds to knowledge. Roush's is used to support her claims that more probable evidence is "ideal" and if there is to be a conflict at all, this must involve evidence in the same sense. To clarify the dispute we shall first consider what Roush's analysis looks like in a "Bayesian" framework, then what it looks like in a "frequentist" framework. By a "Bayesian framework" we mean one in which probabilities represent degrees of belief and in which on learning a new fact e, probabilities are "updated" by changing from the "prior" probability (labelled $P_i(\bullet)$) to a new "posterior" probability (labelled $P_f(\bullet)$) by the rule: $P_f(\bullet) = P_i(\bullet|e)$.

Interpreted in a Bayesian way, Roush's analysis can then be expressed as follows. If the agent holds a higher prior degree of belief in e, but the same values for $P_i(e|h)$ and $P_i(e|\neg h)$, then on learning e the agent would have a higher posterior degree of belief in h than would have been the case had the agent learned e while holding the lower prior belief in e. Though this is a consistent Bayesian account of how higher priors in e can be advantageous for obtaining a higher posterior in h on learning e, there is problem. The agent can only have the higher prior in e, given the other fixed conditional probabilities, if the agent also has a higher prior in the hypothesis. This follows from the fact that the higher $P_i(e)$ also implies a higher $P_i(h)$ given the factors Roush holds fixed. There is then no reason to attribute the higher posterior in the hypothesis to the higher prior in the evidence (as Roush would like) rather than to the higher prior $P_i(h)$.

In contrast the conventional analysis avoids this difficulty. Here $P_i(h)$ is fixed when comparing the cases where evidence has low and high prior degrees of belief. Thus the higher posterior degree of belief in h once e has been learned *is* attributable to the lower prior degree of belief in the evidence under the conventional view. Indeed, this is just the Bayesian story as to why surprising evidence confirms more: Evidence with a lower prior once learned raises the posterior in the hypothesis more assuming fixed values for the probability for the hypothesis and for the conditional probability of the evidence given the hypothesis.

Unsurprisingly perhaps, given this tension between her analysis and a Bayesian interpretation, Roush rejects the premise that P(h) should be fixed when comparing high and low probability evidence. Instead she takes it that high P(e) should make a difference to P(h) *before* learning that e is the case:

"Your degree of belief in e prior to the conditionalisation is just P(e), so high P(e) is (almost) sufficient for you to take e as evidence for whatever e happens to be positively relevant to, that is, to conditionalize upon it. Roughly, if you are confident of e, then you ought to let your other beliefs feel the appropriate effects of e's truth" (Roush, 2005, 174)

However, for a Bayesian this is irrational, since if P(e) is less than one then this *means* that the agent does not believe e is certain and would not rationally "let [their] other beliefs feel the appropriate effects of e's truth".[14]

Roush, however, takes her unorthodox interpretation of Bayesian updating to be virtuous, since it fits with her view that for e to be evidence for anything else it must itself be probable:

"It seems to me inescapable that in order for the value of P(e) that precedes Bayesian strict conditionalization to justify Bayesian strict conditionalization P(e) must be high" (Roush, 2005, 174).

It is as if Roush supposes that Bayesians have a three-step process. Agents begin with degrees of belief represented by the "antecedent" probability P_i. At the first step they observe e. At the second they decide on this basis that the probability of e should be 1. Because the probability of e is 1 they are justified, at the third step, in changing their degrees of belief to those represented by the "posterior" probability P_f. But of course Bayesians do not take three steps, only two. They observe e at the first step and at the second, revise their probabilities in one fell swoop to P_f, which among other features sets the probability of e to 1. For the Bayesian the new probability is justified by learning e, not by the fact that one has become confident of e (ie, already set the probability of e high). The posterior probability is an expression of one's confidence, not a justification of it. The Bayesian is far more objective here than Roush would have it: It is observations that justify new degrees of belief, not simply one's antecedent degrees of confidence.

These difficulties suggest a possible fix: Do not assume e is certain when it is not, but use Jeffrey conditionalization, under which $P_f(\bullet) = P_i(\bullet|e)P_f(e) + P_i(\bullet|\neg e)P_f(\neg e)$. With this approach one might be able to argue that it is better to have a higher prior in the evidence, assuming identical $P_i(e|h)$ and $P_i(e|\neg h)$, than otherwise. However, as it happens this runs into the same difficulty as the strict conditionalization case, namely a higher prior in the hypothesis is implied when there is a higher prior in the evidence. Thus even with Jeffrey conditionalization, the higher posterior cannot be attributed to the higher prior in the evidence.[15]

Perhaps a resolution lies elsewhere. In her examples, [16]Roush describes scientists

[14] Note that Jeffrey conditionalization does not help here, since conditionalization on the original value of P(e) does not lead to any change in the degrees of belief, and conditionalizing on a different value of P(e) is inconsistent with the agent's degree of belief in e.

[15] This is as one would expect given that it is a generalisation of the strict Bayesian case already discussed.

[16] Such as the Rutherford example (Roush, 2005, 174) and her hypothetical medical test example (Roush, 2005, 171).

or doctors finding out that the evidence is probable, arguing that having a high probability here is important. This may suggest that the way to make sense of the importance of probable evidence for Roush is take evidence to be probable *once the agent has become confident of it*, that is, to take her demand that the probability of the evidence to be high to refer to the posterior rather than prior probability of the evidence. [17]This would assume Jeffrey conditionalization, since in the strict updating case the posterior of the evidence is always one, so more or less probable evidence cannot be modelled.[18]

At first blush, this modified approach looks promising. To see why, assume identical priors in the evidence and in the hypothesis in order to allow one to attribute the greater confirmation power to the higher probability of evidence. With Jeffrey conditionalization, the higher the posterior in the evidence, the higher the posterior in the hypothesis.[19] Yet this is not consistent with Roush's analysis, since under these assumed conditions, $P_i(h|e)$ (which under Jeffrey conditionalization equals $P_f(h|e)$) must be the same across the comparison of higher and lower probability evidence. Moreover, this proposed analysis amounts to an argument that it is better to learn more probable evidence because it raises the posterior in the hypothesis more. But since P(h|e) is unchanged, this is not an argument that probable evidence makes for better evidence in a relevance sense. It is rather an argument that learning more probable evidence makes for better grounds-for-knowledge evidence since it leads to a higher posterior in h. But that more probable evidence makes for better grounds for knowledge is not in dispute.[20]

Turn finally to the frequency perspective on probability, from which Roush's argument for probable evidence *can* be made sense of. Consider two populations where event-types h and e both occur, where P(e|h) and P(e|¬h) are the same across both populations, and where LR>1. It follows by Roush's analysis that if e is more long-run frequent – in this sense more probable – in the first population, then the probability that h occurs in the subpopulation where e occurs must also be greater for that population. This shows that if an event-type is more probable (in the frequency sense), it must also be more positively relevant to h, assuming P(e|h) and P(e|¬h) are the same (again in the frequency sense).

Note that by contrast with the Bayesian cases above, the fact that h must also be more probable in the population where e is more probable is not a problem. Far from it, h's being more probable in the first population shows that, in addition to more probable evidence making for better relevance evidence (in that P(h|e) is higher), it makes for better grounds for knowledge in that P(h) is higher. We must be careful, however, about what claim "h" represents. That P(h) is higher in one population than another gives better grounds for knowledge of the claim that a randomly drawn member of that population will be an h. Likewise the other probabilities, P(e) and conditional probabilities, must be interpreted accordingly for the same population.[21]

Finally, it is important to emphasise that not all of Roush's examples of hypotheses can be construed as event-types occurring in (ideally) "infinite" populations. Nor can most hypotheses for which we wish to have a theory of evidence. There are notorious and well-rehearsed difficulties in applying this kind of frequentist account to hypotheses of arbitrary form, which we need not repeat here.

3.3 What we conclude about Roush's defences of high P(e)

In conclusion, independently of her arguments about tracking knowledge (which we turn to in section 5), Roush defends her claim that P(e) be high in order for e to be evidence for anything on three fronts. The first involves arguments based on formula ***A.*** and the accompanying graphs. These, we have argued, provide weak grounds for the demand, if any at all. On the second front she attempts to defuse arguments to the opposite conclusion, that P(e) should be low. One of her arguments on this front is that Bayesians need high P(e) to warrant updating degrees of belief. We have countered that this argument rests on a mistake about the nature of Bayesian conditionalization. In a further attempt to reconcile Roush's arguments for probable evidence with Bayesianism, we also considered Jeffrey conditionalization. The first case considered, Jeffrey conditionalization with a higher prior in the evidence, entailed a higher prior in the hypothesis undermining the attribution the greater confirmation power to the higher probability of the evidence. The second case, Jeffrey conditionalization with a higher posterior in the evidence, entailed a higher posterior in the hypothesis with P(h|e) unchanged, thus showing how learning more probable evidence can be beneficial in a grounds-for-knowledge rather than relevance sense.

[17] Such a reading also fits well with some of Roush's comments: "P(e) reports actual degree of belief, not how much you expected at some prior stage that you would believe e at this stage" (Roush, 2005, 175).

[18] Yet another option would be to move to Jeffrey conditionalization, take P(e) to be a posterior but stick to something akin to Roush's three-step updating by allowing the updating to P(e) before updating other degrees of belief on e. This might better describe how beliefs change in practice and could also help modelling "old evidence" situations, since old evidence could be modelled as evidence which has already been updated upon specifically, but which has not been updated upon generally. However, this approach also requires that the agent hold incoherent degrees of belief which undesirable when one is arguing for a normative model of belief, as Roush is doing.

[19] This follows immediately from the formula for Jeffrey conditionalization: ie, the greater Pf(e) is (assuming LRi>1), the greater Pf(h) will be.

[20] A final attempt to find a suitable Bayesian interpretation of Roush's argument might be to try to show that a higher posterior in the evidence after Jeffrey conditionalization makes for higher Pf(h|e) assuming identical posteriors for Pf(e|h), Pf(e|¬h) such that LRf>1 in line with Roush's formal analysis. We then know, from Roush, that Pf(h|e) must be greater when Pf(e) is greater. But this also fails to give a plausible Bayesian argument for probable evidence since it implies a higher prior in the hypothesis for the case where one learns the more probable evidence (see theorem 4 in appendix).

[21] The importance of taking care when interpreting the probabilities can be seen in Eric Barnes' recent criticism of Roush's medical example (Roush, 2005, 171) where he argues that she equivocates in her interpretation of P(e). See Barnes (2008, 554-555) for details.

In addition to these two fronts, Roush points out that with LR>1 as the criterion, as opposed to P(h|e) > P(h), e can still discriminate even if it has probability very close to 1. She also offers an alternative interpretation to some examples of Peter Achinstein that were supposed to provide cases where "it is the very improbability of e that makes it evidence for h" (p.176). All these show either that P(e) need not be small or that it is no harm for it to be big. This is in line with the view that follows from ***DC*** and ***IC***, that the probability of e is irrelevant to whether e is evidence or not.

The only positive argument for evidence – in a relevance sense – to be probable was revealed when we considered a frequentist interpretation of Roush's argument. Here it was shown that higher probability evidence can be beneficial both in a grounds-for-knowledge and relevance sense. This positive result, however, is somewhat tempered by the well-known limitations of frequentism for assigning probabilities to hypotheses.

4. Leveraging

We want to use evidence for h to arrive at an assessment of P(h) – that's what evidence is supposed to be good for. If the ultimate aim is to use the evidence to arrive at an assessment of P(h), it should not be necessary to assign a value to P(h) in order either to assess P(e) or to assess whether e is evidence for h. In Roush's terminology, we should *leverage* to P(h) by using information other than P(h).

For Roush, evidence must meet both ***DC*** and ***IC.*** To assess if ***DC*** is met, it is sufficient to assess LR and the usual and most immediate way to do this is to assess P(e|h) and P(e|¬h). Formula ***A*** shows that all that is required in addition to assess whether ***IC*** is met is an assessment of P(e). And the discussion following it shows that high enough P(e) ensures ***IC***, given that ***DC*** is met. So an ideal way to satisfy ***DC*** and ***IC*** is for P(e) to be high. If we know this, we can know that e is evidence for h without having to assess P(h).

Leveraging is an idea we entirely endorse [cf. Cartwright (2009)]. Indeed the importance of leveraging cannot be stressed enough when it comes to considerations of the *use* of evidence, considerations that we think philosophers need to keep centrally in view in developing accounts of what evidence "really is". Both pure science and policy want to use evidence for h to help to arrive at a reliable estimate of P(h). This gives yet another argument, based on the idea of leveraging, for concentrating as Roush does on the size of P(e).

Suppose one has gone down the route of demanding that evidence must have a high likelihood ratio, as she supposes, or as many others suppose, must satisfy a relevance requirement, like P(h|e) > P(h|¬e). In both cases, once it is known that either of these requirements is met by knowing the relevant conditional probabilities, it remains only to learn P(e) to fix P(h) because both the following formulae are true:

$$P(h) = P(h|e)P(e) + P(h|\neg e)(1-P(e))$$

$$P(h) = \frac{P(e)-P(e|\neg h)}{P(e|h)-P(e|\neg h)}$$

So from the point of view of leveraging, if one must know P(e) in order for e to be usable as evidence (as one must given Roush's requirement that P(e) is high), demanding knowledge either of the components of the likelihood ratio or the components of the relevance difference is enough. No additional requirements are needed, such as ***DC***. More may be needed to characterize what it is for e to be "relevant to" the truth of h, that is, for our second sense of evidence. But they are not needed for leveraging, which can make do with far less. This is why we think it is important to prise the two notions apart and to allow different accounts for them
In particular, for leveraging, tacking on requirements from the other senses of evidence can be highly restrictive. When it comes to calculating a given target probability, any kind of information that does the job is as good as any other. It all depends on what probability assignments we already accept or can efficiently find out. The calculus of probability constrains the relations among probabilistic facts, but a large variety of combinations can fix the value for a given target. This suggests that no particular constraints should be put on what probabilistic facts should be counted as evidence when it is evidence as leverage to targeted probabilities that is at stake.[22]

Leveraging has two aspects, of which we have so far discussed only one. For evidence of h to be of genuine use, not only should it help us calculate P(h), but it should also be more accessible than P(h) itself. From this point of view we should like to comment, albeit briefly, on a deep and controversial position that Roush defends: that evidence should be characterized entirely in terms of probability. Roush argues that this should be done in order to avoid introducing concepts in the explication of "evidence" that are even more obscure than "evidence" itself. Indeed, she claims, concepts are often offered in explication of "evidence" themselves generally receive their clearest explication in terms of probabilities. Explanatory relevance is a prime example.

We take issue with this last claim, since it has been argued at length that causation, and thereby causal explanation and thereby explanation in general, cannot be given a purely probabilistic explication.[23] But that is not the issue we would like to point to here. Rather we worry about the fact that probabilities are hard to come by. It is for this reason that we urge that the project of explicating "evidence" should start a big step before the starting position of Roush and others who offer purely probabilistic accounts. For the purposes of both pure science and policy it is standard practice – and a practice we would wish to defend – to first gather the evidence and then to use it to assess various probabilities, eg, P(h), P(h|e), P(h|¬e), P(e|h) or P(e|¬h). And for this we need guidelines about what counts as evidence that are not couched in terms of probabilities. One might think of these guidelines as an approach to our second notion of evidence as a two-place relevance relation. But it also helps with the accessibility aspect of leveraging,

[22] Note too that in this case it is not knowledge of events that is being employed but rather knowledge of their probabilities.
[23] Cf., among works by many authors, Cartwright (1979) and Cartwright (1989).

since the idea is to isolate those kinds of facts that will help in the assessment of the otherwise difficult to reach probabilities. This, however, is a project much in its infancy in the contemporary philosophical literature.

5. Evidence and Tracking

What is the connection between tracking and evidence as Roush defines it, and what role does high P(e) play in it? Very roughly, x tracks y means that x and y are correlated: They both obtain or fail to obtain together. Roush is concerned with the kinds of cases common in philosophy of science in which a subject comes to know h via believing evidence e. For this she maintains that high P(e) is required because

"...in this trajectory for knowing h not only is h true but also b(h) TRACKS b(e), b(e) TRACKS e, and e TRACKS h. Now, if h is true – as it must be for anyone to know it – and e TRACKS h then it is unlikely that e is false." (Roush, 2005, 153) [b(x) = "The agent believes x."]

This involves claims about beliefs whereas the relationship between evidence and hypothesis does not involve belief. As one would expect therefore, the relevant concept of tracking for evidence is different. Following Roush (p.150) it can be formulated as follows

> Evidence e tracks a hypothesis h at level u (< 1) if and only if
>
> ***TR1.*** P(e|h) > u
>
> ***TR2***. P(e|¬h) < 1- u.

For Roush, it is important that evidence tracks the hypothesis because that ensures – provided u is high enough and other tracking relations are met – the desirable epistemic goal that belief in the hypothesis tracks the hypothesis.[24]

Note that one could take ***TR1*** and ***TR2*** to define evidence and call it "tracking evidence". By definition then, evidence would track the hypothesis and it would, following Roush (p.154), meet intuitive indication and discrimination conditions. Tracking evidence is also an example of a relevance concept of evidence. Here whether or not e is tracking evidence for h depends purely on the relationship between e and h, and the probabilities of e and h are no part of the characterization. Moreover, it has the nice property that when e is tracking evidence, then P(e) and P(h) must be close in value. But, whether P(e) and P(h) are high or low is independent of whether e is tracking evidence for h.

Given the attractive properties of the tracking definition, one may wonder why Roush does not adopt it. In short, Roush (n.7, p.160) explains that she does not define evidence in this tracking way because so defined, e being evidence for h does not imply P(h|e) > 0.5.Thus e can be evidence for h and yet fail to provide adequate reason to believe h. Nevertheless Roush does not relinquish the aim that evidence should track the hypothesis. So it is important to ask whether evidence as she defines it meets the tracking requirement.

[24] This follows from the "transitivity enough" property of the tracking relation (Roush, 2005, 151-152).

It is not hard to see that it does not always do so. The probability distribution specified by P(e) = 0.2, P(h) = 0.9, LR = 19 has P(h|e) = 0.994, P(e|h) = 0.22 and P(e|¬h) = 0.012. It is evidence under Roush's definition, but given the low value of P(e|h), e does not track h.

More generally, the relationship between Roush's definition of evidence and tracking can be made clearer using two simple bounds on P(e|h) and P(¬e|h).[25]
(i) 1/LR ≥ P(e|¬h)
(ii) P(e)/P(h) ≥ P(e|h) > P(e)

When LR is sufficiently high, bound (i) implies that P(e|¬h) must be low and thus that ***TR2*** will be met. Likewise, bound (ii) shows that a high P(e) is sufficient for ***TR1***. This fits Roush's analysis well, since high LR and P(e) are shown together to be sufficient for evidence to track. Moreover, the higher LR and P(e) are, the better the tracking will be. However, bound (ii) also shows how tracking can fail when P(e) is significantly smaller than P(h) since then P(e|h) must be small so ***TR2*** fails. This is what happens in the example above.

The relationship between Roush's evidence and tracking suggests another rationale for imposing a lower bound on P(e): to ensure that evidence tracks the hypothesis. However, doing this does not just imply tracking. To see why, recall that tracking evidence is a relevance concept. However, when supplemented with a requirement that P(e) be high, one can infer that P(h) must also be (quite) high in virtue of e tracking h. Therefore, imposing a high lower bound on P(e) ensures e tracks h and thus that P(h) is high,[26] which is what is required for a grounds-for-knowledge concept. So high probability of evidence plays a double role, which arguably leads to a conflation absent in the simple tracking concept of evidence. Probable evidence makes evidence track the hypothesis, a feature characteristic of evidence in the relevance sense, and simultaneously makes evidence a ground for knowledge.

To finish, it is interesting to note that Roush's failure to define evidence so that it implies that evidence tracks a hypothesis need not be a serious problem for her concept of evidence. Tracking evidence is extremely powerful when one has it, since if one knows the evidence is false, then one can be pretty sure the hypothesis is false, and conversely. This, though highly desirable, is rarely met in practice. Often evidence speaks for the truth of a hypothesis when we know it to hold, but when false does not say much for the falsity for the hypothesis. For example, let

h: Jill murdered Jack
e: Jill's fingerprints are on the murder weapon.

In this case, e is intuitively evidence for h. However, e does not track h: Suppose that Jill is a careful, intelligent person and if she

[25] The bounds are derived as follows. First, LR = P(e|h)/P(e|¬h), so P(e|¬h) = P(e|h)/LR. But since P(e|h) ≤ 1, it follows that P(e|¬h) ≤ 1/LR. Second, P(e|h) = P(e&h)/P(h), but P(e&h) ≤ P(e) because e&h => e, so P(e|h) ≤ P(e)/P(h). Finally because LR>1, it follows that P(e|h) > P(e).

[26] When LR>1, 1 ≥ P(e|h)-P(e|¬h) > 0. But given P(h)=[P(e)-P(e|¬h)]/[P(e|h)-P(e|¬h)] it follows that P(h) ≥ P(e) – P(e|¬h). Since high P(e) ensures tracking, P(e|¬h) is low and thus P(h) must be at least almost as high as P(e).

had decided to murder someone she would have used gloves, so P(e|h) is low. Given a plausible probability distribution assignment to e and h here, e would be good evidence for h. So Roush's concept of evidence models this situation well. In contrast, the tracking concept is overly strong and rules the fingerprints out as evidence.

6. Conclusion

In this paper we have clarified Roush's definition of evidence and critically analyzed her arguments that evidence should be probable. Roush's first argument, based on formula ***A*** and the associated graphs, we think is weak. These show that high P(e) is sufficient but not necessary for e to be evidence. As we have shown, both of Roush's criteria for evidence can be met maximally and P(e) take any value at all. Our attempts to reconcile Roush's arguments for probable evidence with Bayesianism at best show that probable evidence makes for better grounds-for-knowledge evidence. In our analysis only a frequentist interpretation supports an argument that more probable evidence makes for better relevance evidence, but this is limited to cases where frequentist probabilities can be applied.

The latter part of our paper discussed the role of leveraging and the relationship between evidence and tracking. We strongly agree with Roush on the importance of leveraging, particularly when using evidence in policy. We have argued, however, that Roush's two criteria for evidence are unduly restrictive from a leveraging view. Almost any constraint on P(e) imposes a constraint on P(h) and thus serves the purpose of leveraging P(h). As regards tracking we showed that evidence as Roush defines it can fail to tracks the hypothesis. If Roush's definition is supplemented with a requirement that evidence be probable then evidence does track the hypothesis. This suggests another reason why Roush may see probable evidence as ideal. However, since the concept of tracking evidence is restrictive this provides weak grounds at best for taking probable evidence as ideal.

In sum it may be the case that high probability is a good thing to require of evidence if evidence is to be a ground for knowledge, but when the aim is to assess more accessible probabilities to leverage to P(h), high P(e) has no special advantage. And when a two-place relevance relation is at stake, we think a convincing argument has not yet been made.

Appendix

Theorem 1

LR > 1 and P(h|e) > a > 0 if and only if there exist x>1, 1 > y > 0 and 0 <z ≤ 1 such that

(i)LR ≥ x
(ii) P(e) > y
(iii) P(e|h) ≤ z
(iv) $\frac{x - z/y}{x - 1} = a > 0$

Proof "if"

First, (i) and x>1 imply LR > 1.

Roush derives the following useful formula from the axioms of probability:

$$P(h|e) = \frac{LR - P(e|h)/P(e)}{LR - 1} \quad \ldots \quad (1)$$

Solving for P(e) yields

$$P(e) = \frac{P(e|h)}{LR\,(1-P(h|e)) + P(h|e)}$$

which with (ii) implies

$$\frac{P(e|h)}{LR\,(1-P(h|e)) + P(h|e)} > y$$

$$\Rightarrow \quad P(e|h) > yLR(1-P(h|e)) + yP(h|e)$$

which with (iii) implies

$$z > yLR(1-P(h|e)) + yP(h|e)$$
$$\Rightarrow \quad [z/y - P(h|e)]/(1-P(h|e)) > LR$$

Given (i), this then implies that

$$[z/y - P(h|e)]/(1-P(h|e)) > x$$
$$\Rightarrow \quad z/y > x(1-P(h|e)) + P(h|e)$$
$$\Rightarrow \quad z/y > x - P(h|e)(x - 1)$$
$$\Rightarrow \quad P(h|e)(x-1) > x - z/y$$
$$\Rightarrow \quad P(h|e) > (x - z/y)/(x -1)$$

Finally, (iv) then implies that

$$P(h|e) > a > 0$$

"only if"

Let z = P(e|h) so (iii) holds. From (1) it follows that

$$P(h|e) = \frac{LR - z/P(e)}{LR - 1} \quad \ldots \quad (2)$$

Define the following function

$$f(p,q) = \frac{p - z/q}{p - 1} \text{ for } LR \geq p > 1,\ P(e) > q > 0.$$

Given the continuity of the right hand side of (2), as p → LR and q → P(e) then G(p,q) → P(h|e). Since P(h|e) > a, it follows by the definition of the limit there exist x and y* such that LR ≥ x > 1 and P(e) > y* > 0 and

$$\frac{x - z/y^*}{x - 1} > a$$

$$\Rightarrow \quad x - z/y^* > a(x-1)$$
$$\Rightarrow \quad x - a(x-1) > z/y^*$$
$$\Rightarrow \quad y^* > z/[x-a(x-1)] \quad \ldots \quad (3)$$

Now define y by y = z/[x-a(x-1)].

Given $x > 1$, $0 < z \leq 1$ and $0 < a < 1$ it follows that $y > 0$, and given (3) $y^* > y$ follows. Since $P(e) > y^*$ it follows that $P(e) > y$.

We have shown that $LR \geq x > 1$ and $P(e) > y > 0$ and $y = z/[(1-a)x + a]$. Solving for a yields

$$\frac{x - z/y}{x - 1} = a$$

The result follows. ■

Corollary 1
Given $LR > 1$, $P(h|e) > a > 0$ if and only if $P(e) > P(e|h)/[(1-a)LR + a] > 0$.

Proof
Let $x = LR$, $z = P(e|h)$ and $y = z/[(1-a)x + a]$. For any LR and $P(e|h)$ (i) and (iii) are met and (iv) is met by definition of y. By the theorem therefore, $P(e) > y$ if and only if $P(h|e) > a > 0$. The result follows from substitution of $P(e|h)/[(1-a)LR + a]$ for y. ■

Corollary 2
Given $LR > 1$, $P(h|e) > ½$ if and only if and $P(e) > 2P(e|h)/(LR + 1)$

Proof
Follows from corollary 1 for $a = ½$. ■

Corollary 3
$LR > 1$ and $P(e) > 1/[(1-a)LR + a] > 0 \Rightarrow P(h|e) > a > 0$

Proof
Since $1 \geq P(e|h)$, $P(e) > 1/[(1-a)LR + a] > 0 \Rightarrow P(e) > P(e|h)/[(1-a)LR + a] > 0$, the result then follows from corollary 1. ■

Corollary 4
$LR > 1$ and $P(e) > 2/(LR + 1) \Rightarrow P(h|e) > ½$

Proof
Follows from corollary 3 for $a = ½$. ■

Theorem 2
Given $LR > 1$, for any $x > 1$ there do not exist a and y such that $P(h|e) > a$ and $P(e) > y \Rightarrow LR > x$.

Proof: Solving (1) for LR yields

$$LR = [P(e|h)/P(e) - P(h|e)]/[1- P(h|e)]$$

The right hand side is continuous in $P(e)$ for any fixed value of $P(e|h)$ and fixed non-unitary value of $P(h|e)$. Given this it follows that as $P(e) \rightarrow P(e|h)$, $LR \rightarrow 1$. Therefore imposing restrictions $P(h|e) > a$ and $P(e) > y$ can not imply $LR > x$ for any given $x > 1$, since one can always find a value of $P(e)$ sufficiently close to $P(e|h)$ such that $x > LR > 1$ by the definition of the limit. ■

Theorem 3
Given $LR > 1$, $P(e|h)$ and $P(e|\neg h)$ fixed, $P(h|\neg e)$ strictly increases with $P(e)$.

Proof:

$$\begin{aligned} P(h|\neg e) &= P(h \,\&\, \neg e)/P(\neg e) \\ &= [P(h) - P(h\&e)]/[1-P(e)] \\ &= [P(h) - P(e)P(h|e)]/[1-P(e)] \end{aligned}$$

But by Bayes theorem, $P(h) = [P(h|e)P(e)]/P(e|h)$ so substituting

$$P(h|\neg e) = P(e)P(h|e)[\,1/P(e|h) - 1]/[1-P(e)] \quad \ldots \quad (4)$$

All the terms in the numerator increase, strictly increase or stay constant with increasing $P(e)$ given fixed $P(e|h)$, $P(e|\neg h)$, while the denominator strictly decreases. Therefore, $P(h|\neg e)$ is a strictly increasing function of $P(e)$. ■

Theorem 4
Consider two possible posterior situations after updating using Jeffrey conditionalization on e. In one case one updates one's degrees of belief on e to the posterior $P_f(e)$, in the other to $P_f^*(e)$, where $P_f(e) < P_f^*(e)$.

Notation:
Let $P_f(\bullet)$ denote the posteriors obtained by updating on $P_f(e)$.
Let $P_i(\bullet)$ denote the priors before updating on $P_f(e)$.
Let $P_f^*(\bullet)$ denote the posteriors obtained by updating on $P_f^*(e)$.
Let $P_i^*(\bullet)$ denote the priors before updating to $P_f^*(e)$.

If

(a) $P_f(e|h) = P_f^*(e|h)$
(b) $P_f(e|\neg h) = P_f^*(e|\neg h)$
(c) $LR = LR^* > 1$.
(d) $P_i(e) = P_i^*(e)$.

Then $P_i(h) < P_i^*(h)$.

Proof:
General result: in Jeffrey conditionalization $P(h|e)$ and $P(h|\neg e)$ remain unchanged on updating on e.

So $P_i(h|e) = P_f(h|e)$, $P_i(h|\neg e) = P_f(h|\neg e)$, $P_i^*(h|e) = P_f^*(h|e)$ and $P_i^*(h|\neg e) = P_f^*(h|\neg e)$.

By the axioms of probability:
$$P_i(h) = P_i(h|e)P_i(e) + P_i(h|\neg e)P_i(\neg e)$$

and

$$P_i^*(h) = P_i^*(h|e)P_i^*(e) + P_i^*(h|\neg e)P_i^*(\neg e)$$

Substituting it follows that

$$P_i(h) = P_f(h|e)P_i(e) + P_f(h|\neg e)P_i(\neg e)$$

and

$$P_i^*(h) = P_f^*(h|e)P_i^*(e) + P_f^*(h|\neg e)P_i^*(\neg e)$$

And by (d) it follows that

$$P_i^*(h) = P_f^*(h|e)P_i(e) + P_f^*(h|\neg e)P_i(\neg e)$$

Since $P_f(e) < P_f^*(e)$, by Roush's analysis it follows that $P_f(h|e) < P_f^*(h|e)$ and by theorem 3 that $P_f(h|\neg e) < P_f^*(h|\neg e)$. It follows by substitution of these inequalities into the above that $P_i(h) < P_i^*(h)$. ■

References

Barnes, E.C. (2008). 'Evidence and Leverage: Comment on Roush'. *British Journal of Philosophy of Science*, 59, 549-557.

Cartwright, N. (1979). 'Causal laws and Effective Strategies'. *Noûs*, 13, 419-437.

———. (1989). *Nature's Capacities and their Measurement*. (Oxford: Clarendon Press)

———. (2009). What is this thing called efficacy. In C. Mantzavinos (ed.), *Philosophy of the social sciences. Philosophical theory and scientific practices.* Cambridge: Cambridge University Press.

Howson, C. and Urbach, P. (2005). *Scientific Reasoning: The Bayesian Approach.* (Chicago and La Salle, Ill.: Open Court Publishing)

Roush, S. (2005). *Tracking Truth: Knowledge, Evidence and Science*. (Oxford: Clarendon Press).

Williamson, T. (2000). *Knowledge and Its Limits*. (Oxford: Oxford University Press)

3.2 EVIDENCE, EXTERNAL VALIDITY AND EXPLANATORY RELEVANCE

(From *Philosophy of Science Matters: The Philosophy of Peter Achinstein, 2011)*

1. Introduction[1]

When does one fact speak for another? That is the problem of *evidential relevance*. Peter Achinstein's answer, in brief: Evidential relevance = explanatory relevance.[2] My own recent work investigates evidence for effectiveness predictions, which are at the core of the currently heavily mandated evidence-based policy and practice (EBPP): predictions of the form "Policy treatment T implemented as, when and how it would be implemented by us will result in targeted outcome O." RCTs, or randomized controlled trials, for T and O are taken to be the gold standard for evidence for effectiveness predictions. I question this: Not just whether they are gold-standard evidence, but more, How can they be evidence at all? What makes them relevant to the truth of the prediction that T will work for us?

I am going to follow Achinstein's lead here and suppose that evidential relevance = explanatory relevance, where A is explanatorily relevant to B just in case A is an ineliminable part of a correct explanation of B, or the reverse or A is indirectly relevant to B: There is some common fact that is an ineliminable part of correct explanations for A and B. I shall argue:

1. It's not evidence for us without evidence that it's evidence.

2. Evidential relevance is a conditional relation: E is evidence for H conditional on the non-shared factors that fill out explanations for E and H. Finding these involves a *horizontal search*.

3. To get shared explanatory elements we need a *vertical search*, up and down the ladder of abstraction. If we haven't climbed the right ladder in the right way, an RCT may not show what we think it does.

It follows from my discussion that RCTs cannot play anything like the central evidential role for effectiveness predictions that they are standardly awarded in EBPP literature.

I begin with some terminology, some assumptions, and some simplifying procedures. First, the fact that effectiveness predictions are predictions should not put us off an explanatory relevance account. Just suppose that the predictions are true. Then look for explanatory relevance.

Second, I adopt the probabilistic theory of causality. I suppose that for each effect-type at a time t, O^t, and for each time t' before t, there is a set of factors $\{C_1^{t'},\ldots,C_n^{t'}\}$ – the causes at t' of O at t – whose values in

[1] Research for this paper was supported by grants from the British Academy to study evidence for use, the LSE ESRC Centre for Climate Change, Economics and Policy and the associated Grantham Centre and a UK AHRC grant to study evidence related to child welfare policies. I am grateful to all three for financial support and to collaborators on them for intellectual support. I am also grateful to Eileen Munro for help with the child welfare example and to Adam Spray and Ravit Alfandari for help in editing.

[2] The ideas of Peter Achinstein I draw on here are primarily from Achinstein 2001, 1983. But of course also from his long series of works over three decades from Achinstein 1978 onwards.

combination fix the objective chance at t' that O takes value o for any o in its allowed range. A *causal structure*, $CS^{t'}(O^t)$, for O^t is such a set along with the related objective chances for all values of O^t for all combinations of allowed values, $L_j^{t'}$, of the causes in the set: $Prob(O^t = o/L_j^{t'})$. For simplicity I will usually suppress time and other indices and also restrict attention to two-valued variables. So a causal structure looks like this: $CS^{t'}(O^t) = <\{C_1^{t'},\ldots,C_n^{t'}\}, \{Prob(O^t/L_1^{t'}),\ldots,Prob(O^t/L_m^{t'})\}>$.

Third, I follow the EBPP literature and concentrate on the effect size of T for O in a population: Prob(O/T) – Prob(O/-T).[3] Fourth, I restrict attention to predictions about the effects of policies on populations and not on single units. Fifth, I consider only positive relevance since that fits in a simple way within Achinstein's explanatory account. Sixth, I concentrate on cases where E is indirectly relevant to H because these are the most complicated cases. Finally, for simplicity I assume that the evidence claims in question are well-confirmed – we can reasonably take them as true.

2. Relevance is conditional on unshared factors

The relevance relation I focus on is objective: One fact (E) *bears on the truth of* another (H). This relation holds between facts because of the way nature and society operate; it does not depend on our knowledge of this operation. There are corresponding epistemic notions – like our reasoned judgments about what is relevant to what. These do depend on the state of our knowledge and a variety of other factors as well, such as time and resource constraints or level and type of expertise. Objective relevance is important for policy deliberation predictions: Gathering, discovering, and surveying facts are all costly. We'd like to confine our attentions to facts that matter to the truth of the policy prediction.

"Bears on the truth of" can seem hopelessly vague. So there are various well-known attempts to explicate it with more familiar notions. One takes relevance to be some kind of causal relation. That's too narrow. So too are various kinds of probabilistic relations: There just aren't enough of these in the world to account for all the obvious evidential relevance.[4] Moreover, relying on probabilities puts the cart before the horse when it comes to the needs of estimating if a policy will work. Achinstein's explanatory relevance by contrast fits the bill nicely.

Why should explanatory relevance be a good stand-in for the more abstract concept "bears on the truth of"? My answer is a mix of views of Achinstein and of my own. Just as the relevance relation aimed for is objective, so must the explanatory relation be in order to serve as a marker for relevance. "Explanation" as I use it, then, doesn't mean something that has the right form and is proffered as an explanation; it means something that *is* an explanation. There will be many of these, some of them nested, which is why, as I argue in Sections 4 and 5, we need good vertical searches to find the widest scope of evidential relevance a result can have.

Achinstein has been criticized for using explanatory relevance because this concept itself, it is argued, is in need of explication. I disagree that we need an explication for the task at hand.[5] There are a host of different "thick" relations in nature we label "causal" (like pushing, feeding, lapping up, mailing,…). So too there are a host of relations that we lump together under the label "explains" when explanation serves as a guide to "bears on the truth of". The fact that we cannot give an interesting non-circular explication of "explains" as an objective relation does not mean that we cannot recognize it when we see it – Newton's laws explain Kepler's and my taking an aspirin explains my headache getting better. Nor does it mean that we cannot take certain claims to be generally true of it.

There is good reason why the Achinstein slogan should work for EBPP. To start with, a correct explanation is always evidentially relevant to its explanandum and vice versa. The first follows trivially if one adopts a deductive nomological account of explanation since the explanans cannot hold without the explanandum doing so as well. But, even if one follows GEM Anscombe (1993) in maintaining that an explanans can be enough – it can be as full an explanation as nature allows – without the explanandum obtaining, nevertheless the occurrence of the explanans is undoubtedly evidentially relevant to the occurrence of the explanandum. The converse is trivial since "explanation" is meant to be "correct explanation".

Indirect evidence is harder. E is (indirectly) relevant to H if there is a correct explanation for H that shares a common element, *X*, with some correct explanation for E. $X + X_u^E$ correctly explains E and $X + X_u^H$ correctly explains H.[6] E is evidence that X obtains. But obtaining X cannot be part of a correct explanation for H unless X_u^H obtains. If X_u^H is not the case, then X and X_u^H *cannot* be a correct explanation for H – it doesn't matter how well-confirmed X is. The relevance of E's truth to the truth of H flows through X and it can only do that given X_u^H. E's truth is of no matter at all to H's where X_u^H fails.

Suppose your interest is in whether H is true. But you know that X_u^H is false.[7] Would you pay to learn E? No. Or take a stock philosopher's case: You are asked to predict the color of a bird in the river. Is the bevy of observed white swans relevant? It is if "All swans are white" is part of the explanation of both your bird's color and theirs. But if you are told that your bird is certainly not a swan, all those observations of swan color are worthless to you.

[3] See Coe, 2002

[4] For Anchinstein's views on why purely probabilistic characterizations of evidence do not work, see inter alia Achinstein 2004, 1996, 1981.

[5] For Anchinstein's views on this issue see especially Anchinstein 1981.

[6] Subscript "u" marks the unshared elements of the explanations.

[7] I suppose here that X would not figure in any other correct explanation for H were H to obtain.

So: When the topic is evidence for policy predictions, the relevant concept of relevance is a conditional one: The relevance of a fact E that would have a shared explanatory element with H were H to be true is conditional on the obtaining of the unshared portion of the explanation H would have. Moreover, the epistemic probability awarded to E being relevant should be no higher than the epistemic probability that appropriate unshared factors obtain.

3. External validity and the need for horizontal search

An ideal RCT is a study in which the population in the study, ϕ, divides into two groups that are identical with respect to all features casually relevant to the targeted outcome, O, except for the policy treatment, T, and its downstream consequences. Suppose the probability of O is greater in the T group than the –T group. Where can we go from there?

Under the probabilistic theory of causality, the values of a full set of O's causes fix the objective chance that O takes any value in its range. That's what prompts the attention to the conditional probabilities from the causal structure for ϕ, Prob(O/K&T) > P(O/K& –T), where K is an assignment of values to all the members of $CS_{\phi}(O)$ with the exception of T and its downstream effects. Whether T has a positive effect size in ϕ depends on the relative weights in ϕ of subpopulations in which T acts positively and those in which it acts negatively.

A study is said to be externally valid when "the conclusion established in the study holds elsewhere". Consider an ideal RCT for T,O on a large study population ϕ that has a positive result:

> Study Conclusion (*SC:*) Prob(O/T) > Prob (O/–T) in ϕ.

The study has external validity for target population θ if

> Target Conclusion (*TC*): Prob(O/T) > Prob (O/–T) in θ.

(Recall, θ describes the target population supposing the implementation that would in fact occur given the policy in question.)

> When is SC evidence for TC?

Since neither SC explains TC nor the reverse, if SC is to be evidence for TC there must be some shared part in their separate explanations. The explanation for the successful RCT results in ϕ under the probabilistic theory of causality look like this for some specific causal structure, CS(O), and some specific set of causally homogeneous subpopulations from CS(O), K = {…,K_j,… },

> Study Conclusion Explanation (SCE):
>
> SCE1: The causal structure for O of ϕ is CS(O).
>
> SCE2: For K_j in K Prob(O/K_j & T) > Prob(O/K_j & –T) according to CS(O).
>
> SCE3: The possible negative effects of T on O in other subpopulations are not enough to outweigh this increase.[8]

The explanations for the predicted hypothesis TC are the same in form and must refer to the very same causal structure and the very same causally homogeneous subpopulations if there are to be shared factors:

> Target Conclusion Explanation (TCE):
>
> TCE1: The causal structure for O of θ is CS(O).
>
> TCE2: Some member(s), K_j,K_j,,… of K are subpopulation of θ.
>
> TCE3: For these K_j, Prob(O/K_j & T) > Prob(O/K_j & –T) according to CS(O).
>
> TCE4: The possible negative effects of T on O in other subpopulations of θ are not enough to outweigh the increase due to these.

Since most of the claims in both explanations are indexed to the population, the only shared element is the claim that CS(O) implies that Prob(O/K_j & T) > Prob(O/K_j & –T) for the K_j of TCE3. It is this – and only this – one shared explanatory element that makes the RCT result relevant to the policy prediction. But it is shared only supposing that TCE is a correct explanation for the prediction about θ. That is, the RCT is explanatorily relevant, and thus evidentially relevant, only *relative to* the truth of TCE1,2,&4. What then should be required for the RCT to be accepted as evidence? My dictum: It's not evidence for us unless we have evidence that it's evidence. That means having evidence for TCE1,2,&4. And what reasons do we have to accept these?

To start, what supports TC1 – that ϕ has the same causal structure for O as ϕ? Common causal structures are not all that typical. The refurbished Cuisinart Classic 4-slice toaster that I almost bought for £41.46 has a different causal structure than does the Dualit 3-slice stainless steel toaster at £158.03, which has a different structure again from the new Krups expert black and stainless steel toaster at £44.99. Perhaps you think – as many economists and medical RCT advocates seem to – that your two populations are more likely to share causal structure than are the toasters on offer in Oxford. That's fine. But for EBPP you should have good evidence-backed reason for that.

Supposing that the two populations do have a common causal structure, what assures that some of the very subpopulations K_j of ϕ in which Prob(O/K_j & T) > Prob(O/K_j & –T) are subpopulations of θ? The mix of causal factors that obtain shifts all the time, both across situations and across time. Worse, no matter what mix was there before, in implementing policy we all too often alter that mix. Consider the California class-size reduction program. Reduced class sizes did not improve educational outcomes because the program was rolled out over a short time; the need for teachers doubled within a year but the availability of trained teachers did not. Teaching quality went down, offsetting the good influence of class size.[9]

[8] One can express this more formally, but that seems needlessly complicated for our purposes.

[9] Elsewhere I describe this case in terms of capacities. The same kinds of problems arise in both cases.

Finally, why suppose that were T to increase the probability of O in θ as predicted, that would be due to the positive effects in the shared subpopulations rather than in some subpopulations of θ not shared with ϕ?

These questions need answers, and for EBPP, answers reasonably underpinned by empirical and theoretical support. One cannot just plop SC on the table and *say* that it is relevant to TC. Whether it is relevant depends on common explanatory factors, and presuming that common factors obtain requires good evidence. "It can't count as evidence unless there's evidence that it's evidence".

Clearly this dictum can create a regress. That, however, is the human condition. We have to stop somewhere. But it should be somewhere reasonable and defensible. Consider CCTV cameras.[10] Are they working? A glance at the monitor is generally enough to be reasonably certain, despite the fact that in heist movies elaborate techniques are undertaken to make the monitors lie. For relevance, too, we need reasonable and defensible stopping points for the chain of evidence that shows that evidence offerings are evidence. Consider for a moment not the relevance but the credibility of evidence offerings. Where detailed scientific argument and experiment are involved, this is going to be hard for policy analysts and practitioners to judge. That is why institutions like the Cochrane and Campbell Collaborations or the What Works Clearing House have been set up. If they give a study result high marks, it is generally reasonable for a practitioner to take that on faith.[11]

What then of the evidence for the relevance of SC for TC? Sometimes we can assemble some body of facts that are reasonably well attested and that provide good reasons in favor of claims like TCE1,2,&4. But it is hard. And the very cases in which one most wants to perform an RCT are the cases where there will be least evidence that a positive RCT result for the policy treatment is evidence that the policy will work for us. RCTs are touted as gold standard because only they "control for unknowns", for the factors in the causal structure for O that we don't know are there and hence can't check explicitly are distributed the same in the two groups.

So RCTs come into their own when we suspect that a good many factors in the causal structure for the study population are unknown. But then how are we supposed to produce evidence that those very unknown factors are causal factors according to the causal structure for θ? And that θ has some of the same causally homogeneous subpopulations in which T is positive for O as does ϕ? Finally, how do we estimate that in other subpopulations of θ, T won't have enough negative effects to decrease the chance of O there? The very same epistemic gaps that make the RCT the method of choice also make results practically useless for prediction.

The problems discussed in this section demand *horizontal* search. T can increase the probability of O in some mixes of causal factors and not in others; it can even decrease the probability in some while increasing it in others. A positive RCT result is relevant to a policy prediction only relative to assumptions about the mixes of factors operating in the study population and in the target population. To be justified in taking the RCT as evidence we need to gather information about what other factors operate with T in the two populations. That's what I mean by a "horizontal search". To increase the range of relevance of the RCT we also need a "vertical search", which reviews causes across levels of abstraction.

4. External validity and vertical search

The causes in a causal structure can be more or less abstract; and structures involving factors at different levels of abstraction can all obtain at once. "The trajectories of bodies moving on a sphere subject only to inertia are great circles" is true; so too is "The trajectories of bodies moving on a sphere subject only to inertia are geodesics (ie, the shortest distance between two points)". They are equally true because on a sphere, a great circle *is* a geodesic.[12] Generally the higher the level of abstraction of a causal structure, the more widely it is shared across populations. For example, bodies on Euclidean planes subject only to inertia follow geodesics but not great circles. This matters for explanatory relevance.

An easy way to get a grip on how it matters is to consider some examples. The first is from climate-change modeling, where development economists argue that many of the policies that can help alleviate harmful effects of climate change are things that should be done in developing countries anyway. This is the case of the Bangladesh Integrated Nutrition program (BINP) for providing pregnant women with nutritional counseling, with the idea that poor nutrition is not only due to poverty but also to ignorance, for instance to belief in "eating down" during pregnancy. (White 2009) Of course knowledge by itself is not enough, resources are required too, so the counseling was joined by a supplementary feeding program. This is the kind of factor that comes up in a horizontal search.

An analysis by the World Bank's Operations Evaluation Department found no significant impact on infants' nutritional status. This despite the fact that the program had "worked" elsewhere. What went wrong? A number of reasons suggest that the results elsewhere were not evidentially relevant to the success of the policy in Bangladesh. They might have been. It is natural to expect that

[10] See Pawson and Tilley 1997 for a good use of the example of CCTV cameras in parking lots discouraging car theft to argue the need for what I here call "horizontal search" and to show how understanding the mechanism at work can help with that.

[11] I think, however, that negative judgements by these organizations are often made on bad premises. They tend to presume that trusting to pure method is always better than supposing substantive knowledge claims. That, for example, is why RCTs are gold standard and econometric modelling doesn't get a look in. See Cartwright 2007 for more details.

[12] I shall here be relatively cavalier about the metaphysics of properties. I treat abstract features and concrete ones both as real and I treat them as different features even if having one of these (the more concrete feature) is what constitutes having the more abstract one on any occasion. I take it that claims like this can be rendered appropriately, though probably differently, in different metaphysical accounts of properties.

explanations for the results elsewhere and for Bangladesh success would share an important common element: A general principle

> Principle 1: Better nutritional knowledge in mothers plus supplemental feeding improves the nutritional status of their children.

In fact the two populations did not share this principle.

The first reason for the lack of impact, it seems, is that there was "leakage": In Bangladesh the food was often not used as a supplement but as a substitute, with the usual food allocation for that child passing to another member of the family. (Save the Children 2003) The principle "Better nutritional knowledge in mothers plus supplemental feeding improves children's nutrition" was true in the original successful cases but not in Bangladesh. A better candidate for a shared explanatory element is

> Principle 2: Better nutritional knowledge in mothers with *sufficient resources* to use that knowledge improves children's nutrition.

This principle uses concepts at a *higher level of abstraction*. In the successful cases the more concrete description "food supplied by the supplementary feeding program" counted as an instance of the more abstract concept "sufficient resources", but not in Bangladesh. Not getting this straight is a failure of *vertical search*: A failure to identify the right level of abstraction to find common explanatory elements.

A second reason for the lack of positive impact is also a problem with vertical search.

> The program targeted the mothers of young children. But mothers are frequently not the decision makers, and rarely the sole decision makers, with respect to the health and nutrition of their children. For a start, women do not go to market in rural Bangladesh; it is men who do the
>
> shopping. And for women in joint households – meaning they live with their mother-in-law – as a sizeable minority do, then the mother-in-law heads the women's domain. Indeed, project participation rates are significantly lower for women living with their mother-in-law in more conservative parts of the country. (White 2009, 6)

This suggests yet another vertical move to secure a shared principle:

> Principle 3: Better nutritional knowledge results in better nutrition for a child in those who
>
> 1. Have sufficient resources to use that knowledge to improve the child's nutrition,
>
> 2. Control what food is procured with those resources,
>
> 3. Control how food gets dispensed, and
>
> 4. Hold the child's interests as central in performing 2. and 3.

Just as supplementary food did not count as sufficient resources in the BINP, mothers in that program did not in general satisfy the more abstract descriptions in 2. and 3.

The previous successes of the program are relevant to predictions about the BINP only *relative to* the vertical identification of mothers with the abstract descriptions in 2., 3., and 4. But not all of these identifications hold. So the previous successes are not evidentially relevant. For an RCT to be relevant, and to be justifiably taken as such, we need good reasons to back up the claims that the characteristics referred to in study conclusions, which are often fairly concrete, are the same as the characteristics appearing in principles shared across study and target populations, which are often relatively abstract.

Consider another possible example, this from UK child-welfare policy. In many cases a child's care-givers, though not legally compelled, are heavily encouraged, perhaps even badgered, into attending parenting classes. Consider in this context making fathers attend parenting classes.

First, is "father" to be instantiated by "biological father" or, eg, "male partner of the mother who lives in the household with the child", or maybe "male care-giver"? It may well be that the policy would be effective if the male care-givers or men living with the mother are the target but not biological fathers who are neither on site nor care-givers. If so, to focus on "being a father" would be to move to too high a level of abstraction since only the more specific feature, "male care-giver" or "male partner of mother who shares the child's household", enters into a reasonably reliable principle.

On the other hand "compelling father" or "compelling male care-giver" can simultaneously be too concrete. Different cultures in the UK have widely different views about the roles fathers should play in parenting. Compelling fathers to attend classes can fall under the more abstract description, "ensuring care-givers are better informed about ways to help the child", in which case it could be expected to be positively effective for improving the child's welfare. But it may also instantiate the more abstract feature "public humiliation", in which case it could act oppositely. And of course it can fall under both at once. In any case, if the two more abstract features pull in opposite directions, there will be no reliable principle to formulate at the more concrete level involving "fathers". Nor is this pull in opposite directions an unrealistic hypothesis. We know from empirical research that there are varying outcomes associated with compelling/strongly encouraging parents to attend parenting classes and also that these are correlated with varying motivations. (Barlow et al. 2006) Unfortunately we do not yet have sufficient theoretical probing to explain the variation and the correlations.

5. Troubles for vertical search

To secure explanatory relevance in cases like the BINP, it is necessary first to find and defend a shared explanatory principle. This involves finding the right ladder of abstraction

to climb and knowing just when to stop.[13] But a principle can only be shared between study and target if it applies to both. So it is equally necessary to defend that what happens in the study and what is predicted to happen in the target instantiate the abstract concepts in the putatively shared principle.

This is no easy matter since what in the concrete an abstract property consists in often differs dramatically from circumstance to circumstance. This problem arises regularly in economic climate mitigation and adaptation models (and many other economic models as well). Consider studies of how to change American insurance schemes to provide financial incentives for those living in high risk areas, like the chic Florida coast, to make their homes less prone to risk, for instance by changing the roof construction. (Cf., Kunreuther and Michel-Kerjan 2009 plus references therein.) The models often rely on game theory assumptions that rational agents act to maximize their expected utility.. Here we have to worry about misplaced concretization of the abstract feature "utility". The models typically take money to instantiate utility. But there is a good chance that the targeted agents – say rich owners of beach-front residences – will be more moved by the disruption to their domestic arrangements of having builders at work for months than by any contrary financial incentive that could realistically get built into an insurance scheme.

The same problem of context dependence resurfaces when it comes to measurement, where we see a familiar trade-off: Shared principles require higher levels of abstraction; good measurement, lower. For good comparable measurements, we want specific operational procedures that are carried out in the same way each time the measurement is performed. By contract, the methods for measuring an abstract feature generally differ depending on what more concrete features it consists in, which is not the same from case to case.

We are pulled in two directions here. One: Plump for a false universal concretization in order to secure a universal measure. For instance, measure "educational value added" in new British inner-city academies by counting the number of GCSE's passed at a grade of C or better. Or, devise a measurement definition that more correctly captures the abstract feature of interest across its various concrete instantiations. The danger then is that the definition will be so abstract that we don't know what it consists in from situation to situation. For example, what constitutes human flourishing differs dramatically according to individual circumstances and abilities, natural resources, availability of public goods, need, and the like. So the capability approach of Amartya Sen (1985, 1999) proposes as a measure "the number of lives worth living open to the individual". Or, some propose to measure the economic freedom individuals enjoy by the size of their choice sets. Neither of these provides much of a clue about what we are actually to do to assign numbers or ranks to the individuals to be measured.

For EBPP we look to science for advice. Unfortunately when it comes to fixing what constitutes abstract features in the concrete, science offers at best rules of thumb that are highly defeasible. In particular they are beset by what John Perry (2010) dubs "the failure of enrichment": That A consists in M in circumstances C does not imply that A consists in M in circumstances C & C′ for every C′ consistent with C.

The moral particularism literature is rife with examples where A is a moral feature. Stuart Hampshire (2000), for instance, describes telling stories to philosophical audiences. The stories involve a young intellectual French Fascist, a reader of Celine, held by the Free French, whom Hampshire is sent by the British to interrogate. The French will execute the young man; but they tell Hampshire that he can certainly promise the prisoner – falsely – that he will not be executed in exchange for information. Is it acceptable, or even required, for Hampshire to lie to the young man? Hampshire tells the story differently on different occasions. Often the descriptions can be nested, the more detailed descriptions containing the previous, plus more. Depending on how Hampshire tells the story, the audience is in general agreement about what he should do – but the verdict changes as he shifts from level to level. Enrichment fails.
Hampshire's stories involve highly abstract features – *morally acceptable*, *morally required*. Perry's own example involves specific motions that may or may not instantiate his eating a Brussel sprout at his Dewey lecture, depending on the level of detail of the description of the circumstances. So the abstract feature need not be very abstract at all for the failure of enrichment to appear.

Where then can we find help in science either with the problem of settling on the right level of abstraction to find shared explanatory principles or of ascertaining what the abstract features in these principles consist in for both study and target populations? I don't know an answer. But I am sure it takes both theory and local knowledge, neither of which are much in favor in EBPP communities. Without these, scientific studies like RCTs, which are so highly prized for the credibility they confer on their results, will not be explanatorily relevant to the predictions about what will work for us that we need for practice and policy. And I am sure Achinstein is right for these kinds of cases: If explanatory relevance goes, so too goes evidential relevance. Then we have no scientific evidence to bring to bear and evidence-based policy and practice is out the window.

[13] Stopping matters. Increased abstraction generally goes along with increased generality. So the more abstract the principles you embrace, the more so-far-unexplored concrete predictions you are committed to. My own advice has always been: Don't commit to anything more than you need. That is why I have always urged sticking to the numerous more concrete, detailed laws that explain – and explain in proper detail – the various natural and experimental results we observe rather than committing to the super-abstract laws of high theory.

References

Achinstein, Peter. 2001. *The Book of Evidence*. New York: Oxford University Press.

———. 1981. 'Can There Be a Model of Explanation?' *Theory and Decision* 13: 201-27.

———. 2004. 'A Challenge to Positive Relevance Theorists: Reply to Roush.' *Philosophy of Science* 71: 521-24.

———. 1978. 'Concepts of Evidence.' *MIND* LXXXVII: 22-45.

———. 1983. *The Nature of Explanation*. New York: Oxford University Press.

———. 1981. 'On Evidence: A Reply to Bar-Hillel and Margalit.' *MIND* XC: 108-12.

———. 1996. 'Swimming in Evidence: A Reply to Maher.' *Philosophy of Science* 63: 175-82.

Anscombe, Gertrude Elizabeth Margaret. 1993. 'Causality and Determination.' In *Causation*, edited by Ernest Sosa and Michael Tooley, 88-104. Oxford: Oxford University Press.

Barlow, Jane, Isabelle Johnson, Denise Kendrick, Leon Polnay, and Sarah Steward-Brown. 2006. 'Systematic Review of the Effectiveness of Parenting Programmes in Treating Abusive Parenting.' *Cochrane Database of Systematic Review* 3: 1-20.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

———. 2011. 'The Long Road from RCTs to Effectiveness.' *The Lancet*.

———. 2009. 'What Is This Thing Called Efficacy?' In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, edited by Chrysostomos Mantzavinos, 185-207. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Eileen Munro. 2010. 'The Limitations of Randomized Controlled Trials in Predicting Effectiveness.' *Journal of Evaluation in Clinical Practice* 16: 260-66.

Coe, Robert. 'It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important'. Presented at the Annual Conference of the British Educational Research Association, University of Exeter, September 2002.

Dekkers, Olaf, Erik Von-Elm, Ale Algra, Johannes Romijn, and Jan Vandenbroucke. 2010. 'How to Assess the External Validity of Therapeutic Trials: A Conceptual Approach.' *International Journal of Epidemiology* 39: 89-94.

Embry, Dennis, and Anthony Biglan. 2008. 'Evidence-Based Kernels: Fundamental Units of Behavioral Influence.' *Clinical Child and Family Psychology Review* 11: 75–113.

Hampshire, Stuart. 2000. *Justice Is Conflict*. Princeton: Princeton University.

Kunreuther, Howard, and Erwann Michel-Kerjan. 2009. *At War with the Weather: Managing Large-Scale Risks in a New Era of Catastrophes*. New York: MIT Press.

Mackie, John Leslie. 1980. *Cement of the Universe*. Oxford: Oxford University.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. London: Sage.

Perry, John. 2010. 'Dewey Lecture: Wretched Subterfuge.' American Philosophical Association: Pacific Division.

Save the Children. 2003. *Thin on Ground*. London: Save the Children.

Sen, Amartya. 1985. *Commodities and Capabilities*. Oxford: Oxford University.

———. 1999. *Development as Freedom*. New York: Knopf.

White, Howard. 2009. *Theory-Based Impact Evaluation: Principles and Practice*. New Delhi: The International Initiative for Impact Evaluation (3ie) Working Paper 3.

Research for this paper was supported by grants from the British Academy to study evidence for use, the LSE ESRC Centre for Climate Change, Economics and Policy and the associated Grantham Centre and a UK AHRC grant to study evidence related to child welfare policies. I am grateful to all three for financial support and to collaborators on them for intellectual support. I am also grateful to Eileen Munro for help with the child welfare example and to Adam Spray and Ravit Alfandari for help in editing.

3.3 EVIDENCE, ARGUMENT AND PREDICTION

(Forthcoming in V. Karakostas and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings 2)

1. The Context and the Problem

This paper is about evidence, specifically about evidence for *effectiveness predictions*: predictions that a well-described programme, policy or treatment will work for us, ie, that the programme will result in an improvement in a well-specified outcome if we were to implement it in a targeted situation in a specific way – the way we would in fact implement it. Evidence-based policy advocates have invested a great deal of effort over the last few years in evaluating and providing warehouses for storing what they offer as evidence for hypotheses of this form in various areas of concern, warehouses to be visited by "ordinary" policy makers and analysts. There includes for instance the Cochrane Collaboration for medical studies, the Campbell Collaboration for general social policy, the US Department of Education's What Works Clearinghouse, the George Mason University Centre for issues in criminology and the greater London Authority's new Project Oracle for "Understanding and sharing what really works" against youth violence.

These warehouses advertise that they store programmes that "work" to produce targeted results. We as philosophers know to be wary of sloppy language like that. What they store are programmes for which there is very good evidence that they work somewhere, and, if we are very lucky, in a few somewheres. The warehouse keepers police certain kinds of scientific studies, studies that aim to establish causal connections between a programme and a targeted outcome. Programmes that make it onto the shelves in the warehouse are ones that have been tested in what the warehouse regulations regard as very good studies. In particular the warehouse purchasing rules strongly favour RCT study designs – that is, randomized controlled trials.

What an RCT can evidence directly is that the programme worked there, then, in the study population. What makes that evidence for the effectiveness claim of concern to policy analysts: "It will work here, now, as we would implement it?" What does it take for the RCT result to play a part in a support structure that argues for the truth of the effectiveness prediction? That's my question.

I propose the same answer I urge for claims in any domain where the demands for rigor and explicitness are high, as in warranting conclusions in science or for evidence-based policy, namely what I call *the Argument Theory of Evidence*: conclusions are warranted by good arguments, arguments that are both valid and sound. It is surely trivial to remark that a conclusion is warranted by a good argument. But this reminder helps underline two important facts that are not currently at centre stage in discussions about evidence for effectiveness predictions:

1. Evidence is a 3-place relation: e is evidence for h *relative to* a specific argument A for h. Failing the rest of the premises in A, or relative to a different argument A', the very same fact, e, can be totally irrelevant to the very same hypothesis h.

2. Arguments are like chains: they are only as strong as their weakest premise. Focusing on the argument forces the premises to the fore. Often it is just the ones that aren't generally stated that turn out to be most dicey.

2. The Argument Theory of Evidence

2.1 The Theory and the Reasons for It

What is evidence? More specifically, under what conditions is one empirical claim *e* evidence for a second empirical claim *h*?

I note from the start that evidence is not a natural kind. There is no "correct" theory of what evidence is, as there might be a correct theory of what an electron is. When this is the case our account of what makes for a good theory should be responsive to what needs the theory addresses. The theory I propose started as a theory of "evidence for use", in particular for use in making reliable predictions about what results will be produced by actions we consider taking. The Argument Theory is not confined to this context, however, but should fit anywhere we face the same needs.

A central problem I see everywhere that evidence-based policy is on the tapis is that the way the term "evidence" is usually used lets in far too much. And it does so while at the same time purporting to be very restrictive by subscribing to the highest standards of rigour. In response my theory of evidence is demanding. That is because I agree with a common supposition. It is commonly – and I think reasonably – supposed that

Desideratum
A piece of evidence for a hypothesis should speak for the truth of the hypothesis.

It is with this in mind that I offer stringent criteria. I want criteria such that, once a fact meets those criteria, we should be happy to allow it to weigh in.

Here is what the *Argument Theory* demands of evidence:

A well-established empirical claim *e* is evidence for hypothesis *h* relative to a good argument A (or A,A',A"'…) if and only if e is a premise in A, which is itself a good argument for *h* (or, is a premise in A' which is a good argument for a premise in a good argument A for *h*, etc.), where a good argument has true premises and is deductively valid.

The Argument Theory is akin to Clark Glymour's bootstrapping theory of confirmation (Glymour and Stalker 1980) in which we bootstrap from evidence to hypothesis using background assumptions and inductive logic. On the Argument account, given the other premises – which are like Glymour's background assumptions, h follows deductively from e. For Glymour, by contrast, the conclusion we derive from e and the background assumptions is an instance of h. Then we must use inductive principles to get from "instance of h" to h.

For me h itself is the fixed point that we wish to arrive at. The Argument Theory requires that we do so by a good deductive argument. So I need far stronger background assumptions than Glymour. This, I urge, is all to the good. Science and evidence-based policy gain their high status in large part because they lay claim to being rigorous, public and explicit. These were the demands of Popper and of the Positivists and ones that we should insist on adhering to. There is in principle no objection to inferring h from instances of h in particular cases – so long as it is clear what it is about h and these instances that warrant this inference in this case. Are all the instances the same always? Are they at least all the same in this situation? Does the instance in question have special features that make it characteristic, so that if it holds, h holds? Or…? The Argument Theory demands that the assumption that warrants the inductive leap be explicit in each case so it too can be subject to scrutiny. That's because hiding what it takes for the conclusion genuinely to follow from the evidence is both morally and intellectually culpable in any enterprise that sails under the flags of science or of evidence-based policy.

Two parallel lines of defence support the Argument Theory, one ontological, the other epistemological. That's because evidence is Janus-faced. On the one hand it has to do with truth and truth trackers: with what facts of Nature there are and what other facts can ensure they obtain. I should note that here I take a generous view of what the facts of Nature include. In particular, facts can be expressed by general claims, like Maxwell's equations or the claims of general equilibrium theory in economics, as well as by singular claims, like "The cat is on the mat." On the other hand, evidence has to do with our attempts to arrive at truths: with our hypotheses about what facts obtain and the further hypotheses that provide warrant for them. The two lines of reasoning are obverse sides of the same coin, one expressed – to use Carnap's terminology – in the material mode, the other in the formal mode.

Begin with the material mode. Some facts or sets of facts are sufficient for others: if the first obtains, the second cannot fail to obtain. One fact, f_e, is evidence for a second, f_h, then if f_e is a necessary member of a set of facts sufficient to ensure f_h obtains. Note that "sufficient to ensure X obtains" is not the same as "brings X about". It means just what it says: the one set cannot obtain if the other fails.

If you were brought up in the tradition of Hempel and Nagel you may well be more comfortable with the formal mode version of the parallel lines of defence. Evidence for a claim is supposed to contribute to warrant for the truth of the claim. What contributes to warrant for the truth of a claim are reasons, and what makes some claim e a reason for another h is that e figures in a good argument for h. "Good" here = valid and sound; the premises are true and the conclusion genuinely follows from them. Deduction provides a clear sense to what it means for a conclusion to follow from a set of premises. It is the formal mode counterpart to one set of facts being sufficient in Nature to ensure that a second obtains.

Beware the formal mode though. We are looking for a formal mode counterpart of the relationship in Nature where one set of facts is sufficient for another to obtain. Then evidence can satisfy the Desideratum that a piece of evidence for a hypothesis genuinely speaks for its truth. That is the sense of "warrant" involved in the formal mode account of evidence. Alternatively "warrant for h" sometimes means "justifying a belief in h". That is not the sense at stake here. Belief is an attitude or an action, and, I would argue, there is no context-independent sense of justification for it. Whether it is justified to hold a belief in h depends on what is consequent upon believing it. Will God send me to hell for it? Will I build a bridge supposing h is true, which bridge will fall down if h is false? Will I teach it to my graduate students who might then win a Nobel Prize by taking it as the basis for their research or alternatively, fail to get their PhD because their research went nowhere. Still what I think about justifying belief is an aside since belief is irrelevant to my topic.

Evidence in the sense supposed in the evidence-based policy literature and in the sense required for establishing scientific hypotheses has nothing to do with belief. It has to do with the truth of empirical claims and with what facts ensure that truth. So inductive logics and subjective probabilities have no place in the characterization of evidence for these purposes. Of course they may, if you believe in them, play a legitimate role when it comes to our estimates of whether one claim is evidence for another.

The demand that an evidence claim figure in a good argument – both valid and sound – may seem excessively strong. I actually make a stronger demand. Not only should there be a good argument from e to h if e is evidence for h, but we should not count e as evidence until that argument is displayed. I sometime express this in the slogan "It's not evidence till there's evidence it's evidence." C.G. Hempel's account of explanation also demanded validity and it also majored on deductive arguments. Hempel though allowed that many good explanations in science are enthymematic, in particular they are often not completely laid out. When it comes to evidence for scientific claims or policy predictions, I think it can be a bad mistake to allow this. Both science and evidence-based policy get their status in part from their claims to rigor. As a way to ensure rigor nothing beats laying out the arguments and looking to see how good they actually are.

2.2 Some Objections and Answers

There are a few objections philosophers may have right away to the Argument Theory of evidence. None, I urge, undermines the account.

- On this account of evidence we never know that a claim is evidence because that would require knowing that the claim is a necessary part of a good argument. To know the argument is good you need warrant for the other premises. To warrant those premises you need good arguments; to warrant that these arguments are good you need warrant for the premises in them. Etc, etc. That does not seem to me a problem: it's what good honest evaluation requires. Of course we stop

somewhere; we have to. In the best of cases we stop with claims that can be taken as well established. To the extent that our stopping points are not ones we can take for granted, to that extent we should be cautious about our supposition that a proffered evidence claim really is evidence after all. We know from Otto Neurath that in reasoning we are like sailors who must repair our boats at sea without ever putting in to dry dock to build from firm foundations. So we will always have to trust to some claims we take as true, at least for the nonce. But we should not make our situation worse by neglecting our arguments: without laying out all the premises in all the arguments we don't know how leaky our boat is.

- The Argument Theory implies a number of what might be thought oddities, to all of which I have the same answer. Yes these facts are indeed evidence but they are not usually very useful pieces of evidence for us.

  – Anything true is evidence for a logical truth since anything – any claim at all – is a premise in a good argument for a logical truth. Yes, and so, I maintain, it should be. Anything is evidence for a logical truth. Still I wouldn't advise spending much to buy information about other facts to warrant a logical truth. If you know a claim is a logical truth you don't need to buy information about other facts to warrant the claim. And if you don't know that the claim is a logical truth, you will have trouble warranting that the claim is implied by the fact you buy. Still, if I don't know h is a logical truth but I am assured that if e then h, then e is surely worth learning.

  – *A&B* is evidence for A; *A > B & A* is evidence for B; etc. Yes, they are. But we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here in evaluating the warrant for the conclusion

  – Everything is evidence for itself. That's ok. Any claim does speak for itself. Again though, we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here either.

- The Argument Theory employs a flawed theory of relevance. It lets in as evidentially relevant just the kinds of things philosophers have been at pains to rule out. Consider the canonical example: "John Jones takes birth control pills." Surely this is not evidence for his non-pregnancy. But I think, to the contrary, that it is excellent evidence:

  1. Nobody who takes birth control pills gets pregnant.
  2. John Jones takes birth control pills.

  Therefore: John Jones does not get pregnant.

  Given 1., 2. speaks – and speaks compellingly – for the truth of the claim that John Jones does not get pregnant. What better basis could the truth of this claim have? To suppose that John Jones's taking birth control pills is not evidence for his failure to get pregnant is to confuse the task of providing evidence that a fact obtains with the task of explaining why it obtains.

- If all arguments are deductive then on the Argument Theory

  – There can't be both evidence for a claim and evidence for its negation since evidence claims must be able to participate in good deductive arguments and there can't be good deductive arguments for a hypothesis and its negation. That's okay too. It can still be reasonable to say "We have evidence for h and evidence for not-h" when there are results that can figure in plausible arguments for h and results that can figure in plausible arguments for its opposite. What matters is that we recognize that the results only count as evidence relative to some good argument so that we don't just let the result weigh in without commitment to the existence of these arguments. Of course if there are good arguments that are not deductive and hence the truth of the premises does not guarantee the truth of the conclusion, then it can be literally true that there is evidence for both h and evidence for not-h on the Argument Theory of evidence. But that is as it should be.

  – There can be no evidence for false claims. As soon as you know there is evidence for h by lights of the Argument Theory, you know that h is true. But that seems to me no problem. The problem is coming to know that e is evidence for h. This is a serious job and one of my concerns with the evidence-based policy literature, as I shall explain tomorrow, is that it does not take the job seriously enough, while all the while boasting that that is just what it does.

    Although I don't think our ordinary locutions count for much in efforts like mine here to make precise an everyday concept like evidence so that it can serve specific scientific purposes, I'll just note that often we do use the term "evidence" in a way that supposes that there's no evidence for false claims. If I am accused of cooking the books or murdering Ackerly, I might very well respond, "But you couldn't have evidence for that. I didn't do it."

2.3 An Alternative Account of Objective Evidence and Why I Do Not Adopt It

My insistence that in science and policy we want a sense of evidence in which evidence for a hypothesis speaks for its truth echoes views of Sherrilyn Roush, who has done a great of very instructive thinking about evidence. In her book *Tracking Truth* (Roush 2005), Roush links a theory of evidence with

her theory of knowledge – where the latter has to do with what we are entitled to claim for ourselves as knowledge. Her very first sentence in the chapter "What is Evidence?..." is on the knowledge side: "It is a truism that the better one's evidence for a claim p the more likely one is to have knowledge that p." [p 149]. But like me Roush is keen to keep the enterprises of theory of knowledge and theory of evidence separate:

> ...the notions of evidence that I am aiming for are objective in the following sense. That e is evidence for h is understood as holding in virtue of a factual relation between the statement e's being true and the statement h's being true, not in virtue of anyone's believing that this relation exists. [p 156]

Her basic idea is this: "Intuitively, good evidence for a hypothesis is a discriminating indicator of the truth of the hypothesis," [p 154] where "discriminating indicator" means some appropriate probabilistic analogue of "h is true if e is true and false if e is false".

Formally Roush's account of evidence requires that for good evidence:

- P(h/e) be high.

In order to satisfy what she calls the *leverage condition*, Roush in addition requires

- the likelihood ratio [P(e/h)/P(E/-h)] be greater than 1

Moreover it is highly desirable that

- P(e) be high.

There are a number of reasons that I do not adopt Roush's account, hinging primarily on the fact that it is still too much of a hybrid between a theory of evidence and an account of how to justify our claims to knowledge.

- Where our accounts part company at the start is over what Rush calls "Bayesianism". For her this does not mean a subjective interpretation of probability. Rather – "the Bayesian makes the idealizing assumption that all statements of the language in question possess probabilities. This is in contrast to the approach of classical statistics in which it is denied, for instance, that hypotheses have probabilities." [p 155] For the objective notion of evidence that Roush and I both have in view, though, it cannot be probabilities of statements that matter but rather probabilities of facts. I do not see that there generally are such probabilities. Probabilities for facts arise from chance set-ups, which are a special kind of nomological machine (Cartwright 1999), and while nomological machines are not all that rare, those that count as chance set-ups appear to be a small subset.

Then I disagree with each of her conditions in turn.

- P(e) is high. Roush insists on this in a debate about whether evidence should be surprising, which many, Bayesians especially, require. Her discussion at this point repeatedly refers to degrees of belief despite the fact that she means to be embarked on an objective theory of evidence. And I think that's a clue. If we are thinking about a license to "accept" h, there are a variety of reasons to value observing consequences of h that were not expected beforehand: like worries about accommodation rather than novel prediction, or the demand that h have content that goes beyond summarizing what's already known.

- Notice I say here "expected" – that has to do with subjective probabilities which are not relevant to the objective notion of evidence. On the objective side, I urge that e should be true, not objectively probable. High probability of e only comes in as a demand when we consider whether we should "accept" that e is evidence.

Consider an example where we might all be willing to suppose there are objective probabilities. We have 3 coins:

- For C(1), P(h) = .2
- For C(2), P(h) = 1….it is two-headed.
- For C(3), P(h) = .2

Imagine that the following chance-se-up is in place from time t(1) through time t(3):

- At t(1), flip coin1
- At t(2), if C(1) = h, at t(2) flip C(2)
  At t(2) if C(1) = t, flip C(3)
- At t(3) either h occurs on C(1) or either heads or tails on C(2).

Now consider e = "c(1) = h at t(2)" and h = "heads occurs at t(3)". P(e) = .2. That is low. But e is compelling evidence for h. What I want to underline is that it is compelling evidence not despite its low probability but regardless of its probability. It would be evidence no matter what its probability. Even though e has an objective probability, that objective probability is irrelevant to its status as evidence. This claim is true in general I maintain.

What I would say about e is this: "C(1) = h at t(2)", if true, is evidence not for h but for h' = "At t(2) the objective probability of heads at t(3) is .2". This I think is the right thing to say and it is what follows on the Argument Theory.

- The likelihood ratio is high. This is in aid of leverage. Roush tells us: "…evidence provides leverage on the truth of claims about the world. Specifically, knowing that the evidence statement is true is usually a lot easier than knowing that the hypothesis statement is true, and we use the former to help us make progress on the latter where we could not have made progress directly." [p 158] Damien Fennell and I have elsewhere (Cartwright and Fennell 2009) explained problems we have with thinking the likelihood ratio can do provide leverage in the way Roush wants. I won't rehearse those worries here but rather make a more general point. I don't see how to justify any condition that demands leverage in this sense for an objective notion of evidence. Leverage clearly makes sense when we are in the business of justifying our claims to knowledge or trying to estimate what to expect in the future. Suppose e, if true, is evidence for h. There is no point in spending a lot of money to learn whether e is true or not as an aid to deciding whether h is true when it is a lot cheaper

just to learn h directly. But that has nothing to do with whether e is evidence for h or not.

- P(h/e) is high. Suppose this is so and P is an objective probability and e is true. Then the objective probability of h is P(h) = P(h/e) and on the argument account e is good evidence for this – and that is so whether P(h) is high or not. For Roush it is also evidence for h. One could make this stipulation as part of an objective account of evidence but I think it is misleading. We don't have evidence that h will obtain, just that it can, or might or might well; more precisely, that it has probability P(h) of obtaining. There may be no harm in adding Roush's requirement to the argument account but it will mean that there can be good evidence – in the fully objective sense – for false h's, not just evidence we mistakenly thought was good. Evidence does not provide the same assurance as it does on the basic Argument Theory.

Also, note that if it is added as an allowance on the Argument Theory it would play a different role than in Roush's. For Roush this is what secures the relevance of e to h. On the Argument Theory, that is secured by arguments linking e and h. And that demand should be enforced here as well. We should still demand a good argument – valid and sound – for the claim that P(h) = ϕ.

- There is one other feature on which Roush and I differ but not, I think, disagree. That is on *discrimination*. For Roush e should track h; bracketing issues about probabilities, e should be true iff h is. The Argument Theory requires only that h be true if e is. One could perfectly well add this. "Evidence" even "objective evidence" is not a natural kind with a fixed criteria or a fixed extension. I do not wish to opt for this stronger notion since it is far stronger than what seems supposed in the evidence-based policy literature and in the bulk of scientific cases I am familiar with. In particular it would undercut the claim that positive results in ideal RCTs are evidence for causal claims since positive results imply a causal connection between treatment and outcome but negative results do not show there is none.

- I also have a worry about probabilistic characterizations of evidence like Roush's even when the topic is not objective evidence but rather our entitlement to hold some cognitive attitude to a hypothesis or to use it in some way: probabilistic characterizations put the cart before the horse. Subjective probabilities, at least when we employ them in serious decision making, should have reasons behind them. Like what? Conditional probabilities generally play an important role, like P(h/e). How do we set that? One standard way is look to see if e is evidence for h and how strongly it speaks for h's truth, then set the probability of h given e accordingly. But to do that, we need some independent way of characterizing evidence that does not depend on our subjective probabilities.

3. What Makes RCTs Evidence for Effectiveness?

I have rehearsed the Argument Theory of Evidence because it can provide us with an answer to this question and an answer that matters to getting out predictions right in evidence-based policy.

The current evidence-based policy literature rates positive outcomes in well-conducted randomized controlled trials as gold standard evidence for predictions that the treatment in the trail will work if we implement it in our setting. So, what's the argument?

RCT results are normally *effect sizes: ES* =df the difference in the expectation of the outcome (y) in treatment group and in the control group ($Exp(y)_T$ - $Exp(y)_C$). Causes do not, we suppose, produce their effects willy nilly, at least not here prediction is possible. Rather these effects are generated in accord with causal principles. We can without loss of generality suppose that these principles are of this form:[1]

CP: $y(i) = a + b(i)x(i) + z(i)$

where *y(i)* is the outcome for individual *i* in the population where the principle holds, *x(i)* is the treatment variable, *a* is a constant and *z(i)* represents all the other casual clusters that contribute linearly with x to produce the value of y in i. It is apparent from this principle that x is a genuine contributor to y for at least some individuals i in this setting iff $b(i) \neq 0$ for some i. A well-known argument – which I shall call the *RCT Argument* – shows that, under usual assumptions about ideal RCTs,

$ES = Exp(b)\ (X – X')$

where X = the value of the treatment variable in the treatment group and X', the value in the control group.

RCT Argument
1. $y(i) = a + b(i)x(i) + z(i)$
2. $ES = Exp(y(i)/x(i)=X) – Exp(y(i)/x(i)=X')$
$= Exp(a/x(i) = X) - Exp(a/x(i) = X') +$
$Exp(b(i)/x(i) = X)X – Exp(b(i)/x(i) = X')$
$X' +$
$Exp(z(i)/x(i) = X) - Exp(z(i)/x(i) = X')$
3. x is probabilistically independent of b and w.
Therefore $ES = Exp(b(i))(X – X')$

Premise 3 is supposed to be guaranteed by random assignment of individuals to the treatment and control groups and by masking, quadruple masking if possible. I shall suppose that it holds by definition in an *ideal RCT* and henceforth consider only ideal RCTs. We should remember of course that real RCTs are generally far from the ideal and that randomization only assures the independence assumptions in the long run were the same experiment repeated indefinitely.

So for an ideal RCT, if the effect size is positive, so is Exp(b) which means that b is positive for at least some i. So x is a genuine

[1] The results I shall describe are essentially the same for more complicated functional forms.

contributor to y for some individuals in a population subject to CP. This shows that there is a good argument, *A'*, that has among its premises the evidence claim

> e =df "The effect size of x for y in the population in a well-conducted RCT is ES > 0."

and has as its conclusion

> h_1 =df "x is contributes to the production of y for some individuals in the population in that study."

So e is evidence for h_1 *relative to* the RCT Argument and thereby relative to the other premises in that argument (including especially the assumption that conducting the experiment well – randomizing, masking, etc. –delivered the features an ideal RCT is supposed to have). To establish e's evidential relevance to effectiveness prediction h, we now need to find an argument – a good argument – that I shall call the *Effectiveness Argument*, in which h_1 figures essentially as a premise and h as conclusion.

Before I propose one, I want to point out something about CP, which is often subject to a grave misunderstanding, one that I hope the reader won't have been led into because I was careful with the notation. Often CP is written with the reference to the *i*'s implicit, so it looks like this:

> CP': y = a + bx + z.

In this case it is easy to suppose that b is a constant. But there are few treatment variables x for which this is likely be the case. After all, the treatment is usually only the salient factor, or the factor of focus, in a cluster of factors that together are sufficient to produce a contribution, that is, sufficient *when they all take the right values at once*. To use the terminology of JL Mackie (Mackie 1965), x is cause, yes; but it is an INUS cause of contributions to y: it contributes to y, but only when operating in cooperation with helping factors and often a great many of these. In CP, b(i) represents in one fell swoop the values for i of all the helping factors that are necessary along with x to ensure a contribution to y.

Now to the argument. First we need to formulate a conclusion properly. One version would be

> h_{ES} =df "If x = X were introduced in our setting, as opposed to x = X', keeping fixed all the other causes of y in our situation [except those downstream from x], the effect size would be ES for us too."

So, will x make the same average contribution; that is, is the efficacy, which is measured by the treatment effect in the study situation, the same there as here. Certainly if the same principle holds there as here, *a* will be the same since it is constant. But *b* is not a constant; and the effect size is its expectation – that is, the effect size is an average over x's supporting factors. The average in each situation depends on the distribution of these in that situation. Even if the same principles govern the two, that is no reason to suppose the distributions of support factors would be the same. To the contrary in fact, this distribution very often heavily depends on local circumstances so it is unlikely to be the same.

Anyway, the same distribution is not really what you hope for. What you'd really like is that you have – or can arrange to have – a distribution that favours the good values of b – the ones that provide the largest contribution from the programme. At the least, you will want to have some values for which x's contribution is positive and these should outweigh the effects of those that make x's contribution negative; and if getting negative contributions in some individuals in your setting is to be avoided, then you don't want any of these at all.

Suppose though we can lay aside worries about negative contributions in some individuals. Suppose we want to predict simply

> h_{cont} =df "If x = X were introduced in our setting, as opposed to x = X', keeping fixed all the other causes of y in our setting [except those downstream from x], a positive contribution would result for some members of our population."

What does it take to make ideal RCT evidence relevant? I am going to talk, for short, about whether x *can play a causal role* in the production of y – is it genuinely there in the principle for the production of y for some individuals? Here then is what I take it is the weakest valid argument that uses the results we can get from an RCT there as a premise and concludes that the programme or treatment will contribute positively for some individuals here.

> *Effectiveness Argument*
> 1. x can play a causal role in the principles that govern y's production there.
> 2. x can play a causal role in the production of y here if it does so there.
> 3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.
>
> ---
>
> Therefore, x can play a causal role in the production of y in some individuals here and the support factors necessary for x to make a positive contribution are present for at least some individuals here (ie, x contributes to the production of y for some individuals here).

Where then does the RCT come in? It enters in a different argument, an argument that supports premise 1. That is why I talked earlier about what a study can evidence *directly*. As I use this term, a well-warranted empirical claim *e* is *direct evidence* for a hypothesis *h* iff e figures essentially in a good argument for h – a valid argument with well-warranted premises. Now the RCT Argumnet is a valid argument that takes as premise a positive effect size in an experiment and as conclusion, that the programme contributes to the targeted outcome there in the study situation (post implementation). The other premises in the RCT Argument have to do with further features of the study; for instance

that confounding factors are independent of x. The keepers of the evidence warehouses police these premises for particular studies: they judge how well-warranted the other premises in an argument like the RCT Argument are, mostly on the basis of the study design. So if we find a programme in a conscientious warehouse, we have good reason to think there is a good (valid and sound) argument like A' to warrant the claim that x plays a causal role somewhere – there in the study setting. And that is the first premise in the Effectiveness Argument.

So the RCT result can be evidence for effectiveness here, but it is only *indirect*. It is not a premise in an argument for effectiveness but rather a premise in an argument for a premise. Moreover, its relevance is conditional, highly conditional, since it depends on the validity and the soundness of both the RCT and the Effectiveness Arguments. As in this picture, a positive effect size in an RCT is leveraged into evidence that the program works there (in the RCT setting) by argument the RCT Argument; and "it works there" is leveraged into evidence for "it works here" by the Effectiveness Argument; if either argument fails, the lever drops and evidential relevance disappears with a thud.

Both the RCT and the Effectiveness Arguments are valid, so what really matters is their soundness. We may take it for granted that the RCT Argument is pretty good if we find the programme in a reputable warehouse. What about the Effectiveness Argument? What ensures that its premises are well-warranted? Recall, the two additional premises necessary are:

> 2. x can play a causal role in the production of y here if it does so there.
>
> 3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.

What further arguments support these premises? That's the problem. There are no warehouses for information like this, and the kind of information needed is really hard to come by. I don't see how 2. can be supported without a great deal of theory; so too with 3., in order to identify what the requisite support factors *are*. Then, in addition, 3. will require a good deal of local knowledge to determine if we have here even some of the right values for the support factors, let alone a desirable distribution of them.

Before returning to my overarching message, let me take up two objections to my account of what can count as warrant for an effectiveness prediction beyond the earlier objections to the Argument Theory in general.

First: RCTs are often advocated by people who don't like theory – they think our claims to theoretical knowledge are too slippery; they just don't want to trust to them. That means they don't like my view about how 2. gets warranted. They have an alternative proposal: more and more RCTs, with as much variation in circumstances as possible. I agree that more RCTs, and especially across a variety of circumstances can improve the warrant for an effectiveness prediction. It does so by supporting a premise like 2.: the program plays a causal role here. How? That's the rub. The argument could be by simple enumerative induction: swan 1 is white, swan 2 is white…; x can play a causal role in situation 1, x can play a causal role in situation 2, …

And how good is that argument? For induction we need not only a large and varied inductive base – lots of swans from lots of places; lots of RCTs from different populations. We also need reason to believe the observations are projectable, plus an account of the range across which they project. Electron charge is projectable everywhere – one good experiment is enough to generalise to all electrons; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

I urge that rather than some weak inductive argument, we need a rigorous deductive argument. Then we know just what we are betting on when we bet on the conclusion. So I would add a premise to the effect that x can play the same causal role here as in all those other places, add it so that the challenge is clear: just what is the warrant for this very strong claim. That matters because of the weakest link principle: the conclusion can never have any more warrant than each of its premises individually.

The second objection is this. Surely the best evidence that the program will work here is an RCT here. I agree this would be good evidence – let's not quarrel about "best". *Would be* were it possible. But we never do an RCT here really, here on the same population at the same time. And both matter. A sample is almost never going to be a representative. Representative: that means governed by the same causal principles and having the same probability distribution over the causally relevant factors. And time certainly cannot be ignored. Are the causes the same now as they were when the study was done? That's a particularly pressing question for socioeconomic programme since economists from JS Mill to the distinguished British econometrician David Hendry have worried that past regularities are a poor guide to the future in economics, just because the background arrangement of cause shifts so often, and so unpredictably. Of course the experimental population could be representative enough and the causes at work stable enough. Let's just get this stated explicitly as one of our premises. Then we can think about what warrant there is for these assumptions in our case.

Conclusion

That returns us to my overarching point. Evidence is a 3-place relation. e is evidence for h only relative to some argument or other. That is not a new idea at all, and it may not be very controversial. But taking it seriously matters. It is altogether too easy, when we do not keep the arguments to the fore, to overestimate the warrant that our studies can deliver. The RCT is a good example. It is widely taken in the evidence-based policy literature as gold standard evidence for effectiveness

claims. Though perhaps with a caution. The US Department of Education, for example, warns that trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students. This kind of warning simply conceals what needs to be exposed. What is the argument that makes a particular RCT result evidence for a particular effectiveness prediction? As we have seen, if evidence, it is indirect evidence – there are layers of arguments to get from the study result to the effectiveness conclusion. And they all have additional premises, every one of which, along the way, is essential for the security of the final conclusion. No matter how firm the RCT result is, the effectiveness conclusion – for which it is supposed to be gold standard evidence – can have no greater claim to knowledge than the shakiest of these.

Nor is this unusual. Most of our knowledge claims, even in our securest branches of science, rest on far more premises than we would like to imagine, and far shakier. This recommends a dramatic degree of epistemic modesty. Most of us have adjusted to Neurath's lesson that we are like sailors rebuilding our boat at sea. The conclusions I draw about evidence and the amount of warrant it can confer point to his less familiar warning: the boat is far leakier than we like to think.

References

Cartwright, N, Fennell, D (2009) 'Does Roush show evidence should be probable.' Synthese, 175(3): 289-310.

Cartwright N (1999) The Dappled world: A study of the boundaries of science. (Cambridge: Cambridge University Press.)

Glymour, C, Stalker, D (1980) *Theory and evidence*. (Princeton: Princeton University Press).

Mackie, J L (1965) 'Causes and conditions'. *American Philosophical Quarterly*, 2: 245–64.

Roush,S (2005) *Tracking truth*. (Oxford: Clarendon Press).

SECTION IV

PUTTING EVIDENCE TO WORK

4.1 THE THEORY THAT BACKS UP WHAT WE SAY

(From *Evidence Based Policy: A Practical Guide to Doing it Better,* 2012*)*

Nancy Cartwright and Jeremy Hardie

1. Why Do We Want Theory?
We began I.A with the examples of Bangladesh and California to show how carefully and intelligently chosen policies can fail, even though they are based on excellent evidence that the same policy has worked well elsewhere. The journey from "It worked there" to "It will work here" is not easy. In this chapter we want to set out the theory that lies behind our practical recommendations.

We do so because we want to show that our ideas about deciding policy are principled. Our approach is not just bluff and practical – the plain man's way of cutting through the complexity of statistical evidence, probability theory, and so on to hard, coal-face facts that will lead to realistic predictions. Our advice is rooted in a theory, a theory of *evidence for use*. This is a theory designed specifically for the user's problem of understanding what kinds of knowledge are good for reliable predictions about whether policies will work for you as you would implement them.

For evidence-based policy you need evidence that is both trustworthy and that speaks clearly for or against the policy. There are many sources available to help with the first of these: "When is an evidence claim trustworthy?" Various evidence-ranking schemes tell you how to sort evidence claims that can be trusted, claims that are very well-established, from ones that are more doubtful. Most of these schemes focus on one special kind of claim, that the policy works somewhere. They tell you what kind of study can nail that down – generally with RCTs and meta analyses of RCTs as their gold standard for this – and what kind of studies lend some, but far less, credibility to a claim that the policy worked somewhere. And various policy clearing houses – policy warehouses – will vet policies for you, to ensure that they are well evidenced to work somewhere.

We will not duplicate these efforts here. We are engaged in a different enterprise, one that helps carry you beyond the knowledge that the policy has indeed worked somewhere. Which facts speak for or against the policy working for you – meaning, do or don't lend credibility to it – and under what conditions?

We build our recommendations from a theory of relevance. Relevance matters because knowing the facts is not enough when it comes to assembling evidence. You need to know facts that bear on the truth of the policy prediction. Which facts speak for or against it? Suppose you had an encyclopedia with all the facts about the world in it, forgetting what that could possibly mean, including the facts that you get from RCTs. Which ones should you take note of? The encyclopedia would tell you what is true. It would not tell you what is relevant. Getting the right answers about relevance is important, and ensuring that these answers are well-grounded and defensible is essential if policy is to be evidence-based. That is why we need a theory of relevance.

2. Two Assumptions

Our treatment of relevance for effectiveness predictions is based on two assumptions. One relates to evidence in general; the other is special to the use of evidence that it works somewhere to support predictions that it will work here. It will take some explaining before we can tell you what they are.

2.1 Assumption 1

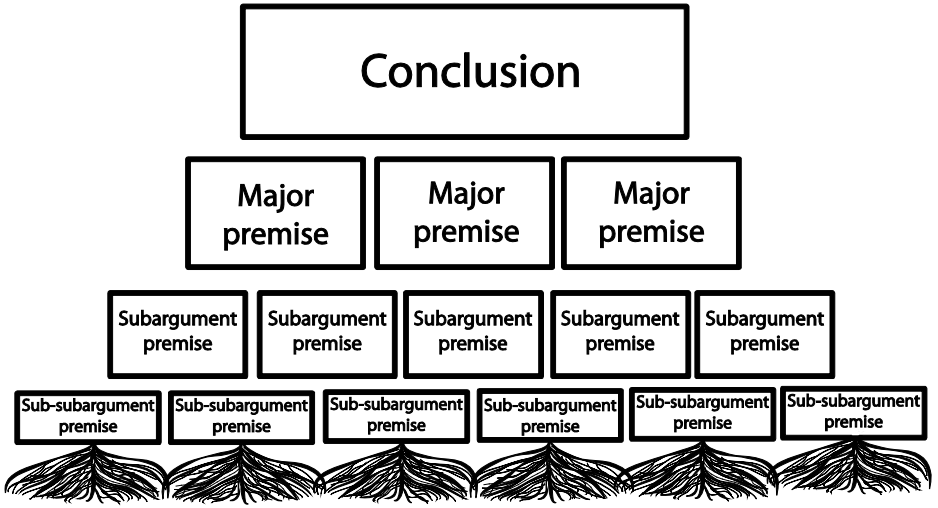
2.1.a What Makes for Warrant?

To warrant a claim is to justify taking it to be true. So, how do you warrant the prediction that your proposed policy will work here, where you are? Warranting a claim, any claim, means marshaling reasons for it so that it is transparent why you have the right to be confident that the claim is true. It follows that *warrant requires a good argument*. An argument here is a set of propositions, called its *premises*, and a proposition, called its *conclusion*, not, as typically in ordinary language, just a reason ("My argument for buying this wine is that it is cheap"). The reasons you marshal must themselves be trustworthy and together they should compel the conclusion, or at least make it likely. That's what we mean by a *good argument*. A good argument is both sound and valid. *Sound* means that the premises are trustworthy; *valid*, that the conclusion is genuinely implied by the premises. Good arguments provide strong warrant for their conclusions; the weaker the argument, the weaker the warrant.

If the conclusion of an argument is to be well-warranted, each of its premises must be well-warranted as well. Some premises may be self-evident, or already well established, or attested to by a reliable expert, or just easy to tell by looking. For instance, from our discussion of CCTV cameras and car crime in III.B.3.1, it does not take a major study to support the claim that your parking lot is surrounded by an 8-foot high wall, nor an intimate knowledge of physiology and criminal sociology to know that most car thieves cannot readily leap 8-foot walls. To know whether CCTV footage is admissible in court, ask a lawyer. And if it happens that a premise in your argument is that a policy has worked somewhere, for warrant you can take the word of a good policy warehouse, like What Works or the Campbell Collaboration. Often, however, the premises in your main argument – call these *major premises* – will themselves need serious support. So you will need subarguments, each with its own premises, to support the major premises. Maybe you will even need subarguments to the subarguments. Each argument and each subargument must be a good one – valid and sound – or the whole structure is threatened.

Sometimes a variety of different arguments can be offered in support of the same conclusion. This is typical when premises are insecure. You hope that at least one of your arguments can stand firm. What's doing the justificatory work in this case is yet another argument with yet further premises. It look like this: "The premises of Argument 1, if true, make the conclusion probable; the premises of Argument 2, if true, make the conclusion probable; …. the premises of Argument N, if true, make the conclusion probable. Probably the premises of at least one of Arguments 1, 2, …, N are true. Therefore the conclusion is probably true." This is fine, so long as you are clear about just what the overall argument is and about how secure its premises are. Otherwise there is danger that you will over – or under – bid your cards.

To get clear on just what your argument is, one device you can employ is to build an argument pyramid, as in Figure I.2. The conclusion is at the top. The major premises are next layer down. Below each major premise are the premises in the subargument that supports it. And below each of those, the premises of the sub-subargument that supports it. Put blank boxes where you can see you need more premises to make a valid argument but don't know just what they are. For instance, Inspector French's argument that Carey killed Ackerly, which we discuss in I.B.2.1.b, might require a motive for the killing. If French suspects there is a good motive, but doesn't have a good hypothesis what it is yet, he would leave the space for that premise blank. Stop when you run out of arguments or don't need them anymore. From any premises that don't need further serious support – like those that are self-evident or that you just see by looking or are already well established or attested to by a reliable expert – draw roots into the ground. This is to show that they can stand on their own.



[Figure I.2]

There are two big advantages to taking seriously the connection between argument and warrant and to sketching argument pyramids like this. First, it helps you put order into your reasons. Second, it helps you assess the degree of confidence you should have in your conclusion. Ideally every box is filled in and every box stands on others and those on others until eventually all are rooted in the ground. Then you can have a high degree of confidence in your conclusion. If lots of boxes are hanging in the air and there are lots of blank boxes as well, your degree of confidence should be low.

Often you will find the pyramid is shaky, but it doesn't seem too implausible to think that the rest, with world enough and time, could be filled in and supplied with thick roots at the bottom. That justifies some confidence, but not a high degree. If you find yourself in this situation – as we suspect you very often will – you should not despair. This is the typical human condition; relative certainty is hard to come by. You can then decide whether the policy is worth pursuing given that it may well be effective, but then again, it may well not. And at least you will have a better estimate of how confident you should be.

There is one caution in the use of argument pyramids. The visual representation can be misleading. Suppose you have an argument with three major premises. If you take away the one that you have pictured in the middle, it can look as if the conclusion is still well supported. But, to the contrary. Where there are blank boxes there are holes in the argument, and an argument with holes provides no support at all. You only get some support if you hypothesize that these holes can be filled in. And your support for the conclusion can be no stronger than your justification for this hypothesis.

2.1.b Evidence and Argument

We now turn to evidence. The evidence for a claim is supposed to help provide warrant for it; it is supposed to help justify your confidence that the claim is true. That means that evidence must figure in a good argument. The evidence claim must appear as a major premise – or a subpremise, or a sub-subpremise – in an argument alongside other premises that together make the conclusion probable. That's what secures relevance. It is the overall argument that turns one of those millions of trustworthy claims one can make about the world into a piece of evidence for the conclusion. This means that whether a particular claim is evidence for a conclusion depends on some specific argument for the conclusion in which that claim figures and on how good that argument is.

Evidential relevance then is a 3-place relation. It involves an evidence claim, a hypothesis (or conclusion), and an argument: Claim *e* is evidence for hypothesis *h relative to some good argument A*. Relative to the argument under consideration, some evidence claims will be what we call *directly relevant*; others only *indirectly relevant.* The claims expressed in the major premises of the argument are what we label *directly relevant* to the hypothesis. But we know that if the hypothesis is to be well-warranted, each of the major premises must itself be well-warranted, so each of these too must have a good argument to support it. Any of the claims offered as premises in one of these subarguments are, in our terminology, *indirectly relevant* to the original hypothesis. This carries on, generating evidence claims that are relevant to the original hypothesis, but more and more indirectly, and relative to a series of arguments connecting them to it.

This gives us a good, succinct account of evidential relevance, that is, an account of what turns a trustworthy claim into a piece of evidence. It is the first assumption of our theory:

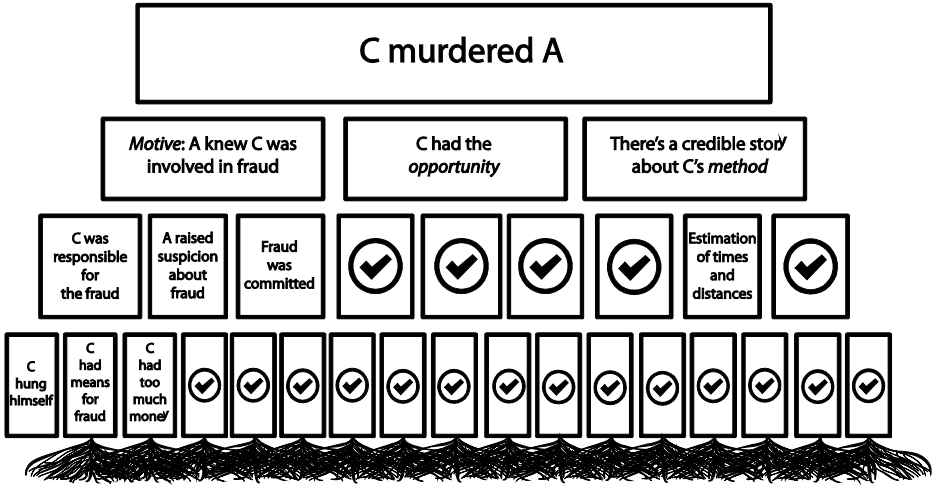
> *Assumption 1*
> A well-established empirical claim *e* is evidence for *h* if and only if *e* can be rendered as a premise in an argument, *A*, and *A* is a good argument for *h*; or, *e* is a premise in *A'*, where *A'* is a good argument for a premise in a good argument, *A*, for *h*; and so forth.

As we urged in the last section, it is important to be clear just what the arguments are for your conclusion and how well supported their premises are in order to assess how confident you can be that the conclusion is correct. An argument pyramid with gaps where you don't know just what form one of the essential premises takes provides shaky support for your conclusion. It's far worse when you have good reason to think one of the essential premises is false. We stress the importance of arguments because arguments, like chains, are only as strong as their weakest link. A valid argument with 9 premises known to be true and one known to be false does not make its conclusion 90 per cent probable. It provides no warrant at all. So a body of evidence that nails down nine of the premises is not enough for high warrant. Indeed, if the 10th premise is known to be false, the argument is no good at all and the proffered "evidence" is irrelevant to the truth of the conclusion. This is important to keep in mind for evidence-based policy, where some of the premises may be very well established but often others may be fairly dicey. In this case you must be careful not to overestimate the warrant you have for your policy predictions.

Mystery stories can provide good examples of the importance of each premise to the stability of the overall structure. Consider Freeman Wills Croft's Inspector French, who is always at pains to lay out clearly the arguments that support his conclusions. In (Crofts 2001), Inspector French has become convinced, on the basis of a body of seemingly good evidence, that Carey murdered Ackerly. We reconstruct his warrant for this in the form we advocate – that of an argument with explicit premises and explicit subpremises.

Premise 1 in French's argument concerns motive. Ackerly, French argued, had cottoned on to a major fraud that Carey was involved in. This was backed by compelling evidence for three claims: fraud had been committed; Ackerly had raised suspicions about it; and Carey was responsible for the fraud. This last was supported by evidence that Carey was in a position to perpetrate the fraud, that he had income otherwise unaccounted for, and, very importantly, that he had committed suicide when it looked as if the fraud might

be revealed despite Ackerly's death.
There was equally good evidence for Premise 2 concerning opportunity. Carey *could* have committed the murder. We won't tell the whole story here: the check marks in Figure I.3 mean that French did indeed have good reasons to back his claim about Carey's opportunity. But opportunity and motive by themselves do not support a conclusion of guilt. French, though, had a third premise. Premise 3 laid out the method. This was a step-by-step narrative, of the kind we discuss in III.B.2, of how Carey was supposed to have carried out the murder. This narrative was supported by good evidence that Carey was at some of the right places at the right times, and French had established, by timing distances and estimating speeds, that Carey could have been at the other places in the chain just when needed. We graph French's reasoning in the argument pyramid picture in Figure I.3. French had a way to fill in all the boxes all the way down, but we won't do so since the details for them do not matter to the story.



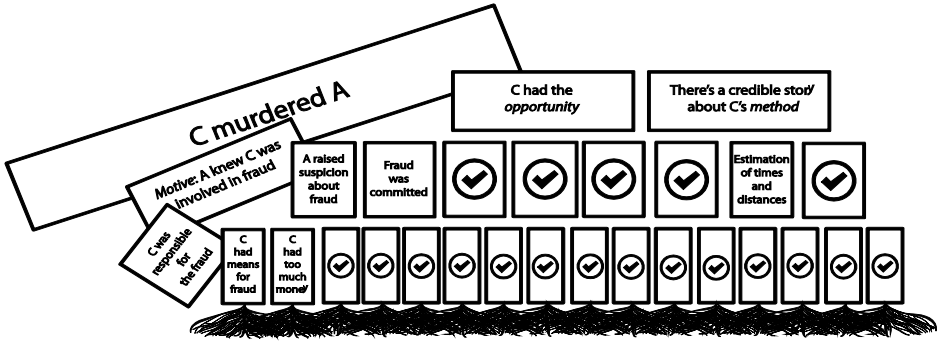
[Figure I.3]

Together French's three major premises made a compelling argument for Carey's guilt – seemingly both sound and valid. Then it was revealed that Carey had not committed suicide but rather was murdered. French immediately reported this to his superior. Here is a record of their conversation, beginning with a question from French's superior (Crofts 2001: 202).

> "This is going to mean an upset to your theory, Inspector?"
>
> French nodded. "Complete. I've got to start from the beginning again."
>
> "Is it as bad as that?"
>
> "…I think so….This murder of Carey makes it unlikely that he killed Ackerly, and if he didn't kill Ackerly, the whole of my theory goes west."

So, the whole theory falls. There is no longer compelling reason to think that Carey murdered Ackerly.

But what about all that evidence that French had so painstakingly gathered? It was no longer evidence that Carey was the killer. The suicide was an essential prop for the assumption that Carey was responsible for the fraud. Without the suicide the other evidence for this was just far too weak to justify that assumption. But without it, Premise 1 fails – motive is not established. And without motive, opportunity and the existence of a possible method for carrying out the murder provide weak warrant, if any at all, that Carey was the murderer. It is still possible that he was guilty. But French no longer has evidence to support that. Evidence is a 3-place relation. The measurements of distances and times are evidence of Carey's guilt, yes. But only relative to French's entire argument. When one of the necessary premises in that argument fails, those measurements are no longer evidence at all.[1] As French says, his whole theory collapses – as in Figure I.4:



[Figure I.4]

[1] Unless of course French could produce another valid argument in which the measurements figure and where all the other premises can be assumed with confidence.

Before turning to our second assumption, some remarks on terminology are in order. Sometimes "evidence" is used in a broader sense than ours. For instance, historian of physics Peter Galison argues that mathematics is the new laboratory of physics. In some branches of physics, like string theory, often the best evidence that a theory is on the right track is not whether it predicts new experimental results but rather whether its equations satisfy certain abstract mathematical constraints, often having to do with symmetries (Galison 2004). When we use the term "evidence" we mean something more narrow, we mean *empirical evidence*. Whether the facts are local, like "There is an 8-foot wall around our parking lot", or very general, like the law of gravitational attraction, for us evidence claims report facts about the world.

Others use the term "evidence" more narrowly than we do. Evidence claims are restricted to reports of the results of individual scientific studies. We reject this usage because our topic is evidence for predicting effectiveness, and the context is evidence-based policy. There is a general assumption about evidence-based policy that if you have a great deal of trustworthy evidence, you are in a good position to predict whether the policy will work here. But if evidence is restricted to just results of individual scientific studies, this is badly mistaken. You need a lot more facts than specific results of specific scientific studies to argue that a policy will work here. You need, as we shall show, facts about causal roles and about support factors—that's what our book is about. Without these facts, and without good warrant for them, your conclusion is not justified. If evidence-based policy is to do its job then, it is best to construe evidence widely enough to cover all the facts without which you will not have a good argument.

2.2 Assumption 2

We begin with the observation that evidential relevance depends on the type of the hypothesis. Different types of hypotheses require different types of evidence. The hypotheses that we are concerned with in this book – policy effectiveness predictions – are causal claims: The proposed policy will *cause* an improvement in the targeted outcome if you implement it. That's part of why you are willing to adopt the policy – because you expect it to make a difference.

The most straightforward argument to support the truth of a prediction is one that lays out the facts that would make that prediction come true. We are going to focus on this kind of argument because it is special: its premises must be true or the prediction won't come true; if any of the premises fail, so too will the conclusion. When the conclusion is a prediction about what effects a policy would have in a given situation, these are the facts about the situation that ensure that the policy would produce the specified effect there. What are these?

We suppose that causes do not produce their effects by accident, at least not if you are to be able to make reliable predictions about what will happen if you intervene. Rather, if a cause produces an effect, it does so because there is a reliable, systematic connection between the two, a connection that is described in a *causal principle*. Our advice is grounded in basic assumptions about these causal principles and how they operate, which we describe below. Facts about these principles will play an essential role in the argument that justifies the prediction that the policy will work here, where you are.

What then about RCTs? How do they figure into our story about the need for a good argument and about the importance of citing facts about causal principles among the premises in arguments for policy effectiveness? After all, RCTs, and related studies that provide strong warrant that the policy worked somewhere, are the conventional focal point of evidence-based policy. If your argument for policy effectiveness does not include evidence of this kind, you will probably be judged not to be doing evidence-based policy. And we agree that starting here can be a good idea, especially now that so many warehouses for vetting and storing this kind of evidence are available. So how do you get RCTs to figure in a good argument for effectiveness?

RCTs show that a policy works somewhere. That's supposed to be one premise. What other premises are needed alongside this to produce a good argument that the policy will work here? This is our second assumption:

Assumption 2

> To get a good argument from "It works somewhere" to "It will work here" facts about causal principles here and there are needed.

Otherwise there is no good way to make the study results relevant to predictions about what will happen were the policy to be implemented in the target setting. Our approach is based on identifying just what facts about causal principles are needed to fill in the missing premises. Our advice throughout is principled in that it is based on an account of causation that is intellectually robust and that translates into practical suggestions for ways to recognize the facts that are relevant to policy prediction.

3. Causal Principles

3.1 What is a Causal Principle?

We take from recent philosophical work on causality a few basic assumptions about the kinds of causal principles that typically support reliable policy prediction.

1. Causal principles do not record mere statistical associations; "Correlation is not causation". That's because causes make their effects happen; they contribute to their production; they are responsible for them. That's the point of adopting a policy. The policy should make a difference. And that should not just be a matter of chance.

2. There may be some causal principles that hold everywhere; perhaps the law of gravitational attraction, that two masses attract each other with a force GMm/r^2, is an example. But this is not typical in the sciences

and especially not in the medical and social sciences. In these areas, principles can vary from one situation to another; they can be more or less general; and more or less deep. In II we shall be reminded how the fact that causal principles can differ from locale to locale means that you cannot read off that a policy will work here from even very solid evidence that it worked somewhere else, or even in a number of somewhere elses.

3. Causal principles are not all on a par. Some may be more fundamental than others. The less fundamental hold *on account of* the more fundamental; with the right additional assumptions about local structure, they can be *derived from* the more fundamental ones. Generally the more fundamental hold more widely. Planets circulate the sun in elliptical orbits that are described by Kepler's laws. But the orbits are also described by Newton's more fundamental laws and, with background assumptions about the starting velocities and masses and their arrangement, Kepler's laws can be derived from Newton's. Kepler's elliptical orbits are the working out of Newton's laws when planets interact with the sun; they are what Newton's laws amount to given the structure and features of the planetary system.

This classic case is from physics. But the same kind of layering of principles occurs in the biological and social realm. In II.B we shall see how this fact can sometimes provide a powerful tool for constructing new, different programs from ones that worked elsewhere or for avoiding failures, as in the Bangladesh Integrated Nutrition Project.

4. The causal principles employed throughout the biological, medical, and social sciences are *ceteris paribus* principles. They hold only "other things being equal", or, to put it more accurately, other things being "normal", or "within appropriate bounds". This is familiar. Reducing class sizes will not help reading scores if a hurricane wipes out the schools.

5. Causal principles can be deterministic, as in classical physics, where the causes fix exactly what effects must occur. They may be merely probabilistic, as in dicing or quantum mechanics, where a given effect occurs with some fixed probability. Or, as in our more general experience, they may be even less regimented than that. The effects occur sometimes, or more often than not, or most of the time. They may also be quantitative, expressed in equations; they can alternatively be qualitative, relating features that do not have exact quantitative measures.

6. As we explain with simple examples in II.A, generally for the kinds of effects aimed for in policy planning

> a. contributions to the effect can come from different sources and via different pathways,
> b. the overall effect depends in some systematic way on the contributions from these separate sources – in the simplest case the contributions simply add up, and
> c. in general social policies on are not enough by themselves to ensure a contribution to the targeted effect; the policy described needs the right *support team* before it can be expected to produce a contribution.

Philosophers have a technical term to summarize 6, which we shall explain in detail in II.A. They say, "Causes are INUS conditions for their effects".[2] An INUS condition is an *I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition for producing a contribution to the effect. We introduce this philosopher's edict because it underlines two important facts to keep in mind in evaluating whether a policy will work here.

First, the program or policy under consideration will seldom be sufficient by itself, no matter how much effort has been taken to include in its description as much as possible of what is required. The policy will only be a *part* of a team of causes that work together. It takes the whole team to produce a contribution. Together, the factors in the team are *S*ufficient for a contribution. But the separate team members, though *N*ecessary for that team to produce a contribution, are each by itself *I*nsufficient for doing so.

The reminder that a policy under consideration is generally just part of a team and is insufficient by itself to produce a contribution focuses attention on the questions: Are the requisite team members – which we call *support factors* – at hand? If not, can they be introduced practicably? There may of course be more than one set of support factors that will round out a policy to form a complete team, in which case the questions is: Are there to hand all of the factors required in at least one team in which the policy figures?

Second, the team in which the policy figures will not generally be the only thing that could contribute to the outcome. The team with the policy in it is *U*nnecessary for a contribution since a number of other teams that may or may not contain the policy can also contribute. The actual value of the outcome in a situation will depend on all the teams that operate in that situation, and on the size and direction of the contribution from each. Some of these will contribute positively and some negatively.

To estimate the *value of the outcome* after the policy, you need to catalog all those factors that will be operating to produce the effect, which is a tall order. Plus you will need some formula for calculating how they combine, how they "add up".[3] You will need to do a little less if you want to predict by *how much matters will be different after the policy*, since for this you need to consider the effects only of teams containing factors that change, but you will need to consider both those that change as a result of your actions and those that change independently of what you do. To predict instead whether and how much things will change *as a result of the policy and its implementation*, you only need

[2] Note that we claim that causes are INUS conditions. But not the converse, that all INUS conditions are causes. The terminology comes from Mackie (1965).

[3] See 7. immediately following.

to know about the teams that include factors that you will be changing – both those that contain the policy itself and also any others that you may change during implementation. We discuss post-implementation effects further in 4. For the more restricted prediction of just *what the policy itself does*, independent of any other changes introduced in implementing it, you need only look at teams with the policy in them.

Even for these final two more typical and somewhat easier exercises, though, it is not always safe to ignore teams that you won't be changing. For instance, if there are already present teams that produce very large contributions, your policy may make such a negligible difference that it is not worth pursuing. There is no point in feeding the prisoner a meal low in salt and fat in order to improve her health if she is to be executed the next morning.

7. There can be different rules for how contributions combine depending on what the subject matter is. The social sciences often assume simple addition. Even there, though, allowance is made for threshold and marginal effects: If the outcome reaches a certain size, additional contributions make no difference, or each new unit may make less difference than the one before. In mechanics, forces combine by vector addition. In econometrics, sometimes the contribution of each separate factor is represented in a different equation. When the factors all act together, the separate equations must all be satisfied at once. And in both physics and economics, it is often supposed that the net outcome will be some kind of equilibrium among the various contributions.

Even though many of the causal principles that ground policy predictions may be purely qualitative, as explained in 5., much of the standard literature supposes that principles can be expressed in equations. We shall focus on principles like this in order to exploit some of the ideas and derivations already available in this literature. For similar reasons we shall focus on equations in which contributions from different sources simply add, despite the more complicated rules for combination noted in 7. We shall also, as is standard, suppose the equations are deterministic, despite the cautions in 5. Without these simplifications, technical matters become more complicated. But the basic lessons we draw remain the same.

Given these simplifying assumptions, we can suppose that the causal principles for the production of an outcome *y* look like this:

CP: $y(i)\ c= a_1 + a_2y_0(i) + a_3b(i)x(i) + a_4z(i)$,[4]

where the *i*'s range over the individual individuals in the population to which the principle applies (these could be anything from individual students to classrooms

[4] We wish to stress, as remarked in the text, that we use this simple linear form to avoid needless complications and to make the discussion easier to follow. Exactly the same lessons follow from more complicated forms, including, especially, the lesson that the RCT treatment effect is a function of the average over the values of the support factors in the RCT population.

or states), *y(i)* is the outcome, *x(i)* is the policy variable, *a*'s are constants across all individuals, $y_0(i)$ is a "base level" of y for *i*, *b(i)* represents all the different factors in all the support teams that work with *x* to ensure a contribution for *i*, and *z(i)* represents, in one fell swoop, all the other factors and their support teams that contribute additively with *x* but do not include *x*. a_2, a_3, and a_4 represent what might be called "boost factors"; they fix the size of the contribution from given values of the variables that follow. (These are like the constant of gravity, *G*, in the force of gravitational attraction between two masses: GMm/r^2.) We employ the symbol c= to represent that the quantities on the two sides are equal and that the ones on the right-hand-side are causes of the one on the left.

3.2 An Illustration

To illustrate, here is a simplified version of one of the examples that we use in II.A, where we set out the causal cake metaphor as a more user friendly way of approaching the notion of INUS conditions. We emphasize later that premature or optimistic simplification is in general risky in thinking about social policy – much of the problem is that you can forget too readily how many and how various are the other factors that have to be present if your policy is to work. So the simplified version here must be seen as no more than an attempt to clarify the notion of an INUS condition, which is a long way from the complexities of a discussion of what actually has to be done when you are to make that notion operative in the field.

Consider the case of Tamil Nadu, where infant health (i.h.) was improved by educating mothers about nutrition. We shall take this result as given. How can that be expressed in terms of our equation CP? Like this:

TN: $\text{i.h.}(i)\ c= a_1 + a_2\text{i.h.}_0(i) + a_3b_m(i)\ e_m(i) + a_4z(i)$,

where e_m is educating the mother, and we know what that means. But the other factors are unknowns.

Suppose, just so that you can see what is at stake, the effectiveness of e_m in Tamil Nadu has been very well established via a very good RCT. A positive effect in the RCT in Tamil Nadu tells you two things: that e_m is a member of some team of factors, b_m and e_m, that produced a positive contribution to infant health for some individuals under the causal principle that obtained in Tamil Nadu; and that the support factors for e_m required under that principle, represented by b_m (whatever they may be), were present in Tamil Nadu. In the terminology we shall employ, e_m *played a positive causal role* for some individuals in Tamil Nadu. This means no more than what the RCT tells you. It does not tell you what causal principle was operating there, nor anything about the other factors that affected the outcomes in the individuals in the study (represented by *z*), nor anything about the support factors represented in b_m, except that they were present, nor about the boost factor a_2.

You now go to Bangladesh and educate mothers in nutrition, betting on an equation like TN:

BD: i.h.(i) c= a_1' + a_2' i.h.$_0$(i) + a_3'b_m(i)'e_m(i) + a_4'z'.

We write a_1', etc. rather than a_1, etc. because here, too, you do not know what a_1', etc. stand for. But whatever they stand for, there is no reason to think that it is the same as in TN. Nor, more important, do you know whether b_m', which represents the support factors for e_m, is the same as b_m.

Post hoc evaluations indicate that the program had little effect on infant health in Bangladesh. There are then two possibilities:

1. That e_m does indeed form part of the applicable causal principle BD, but the support factors were not present.

2. That e_m does not even appear in the causal principle, so the question of the presence or
absence of support factors, such as b_m and b_m', does not arise.
In cases like 1., we will use language like "*could* play a positive causal role". It actually *does* play a positive causal role only if the support factors are present as well. In cases like 2., there is no question of either an actual or a potential causal role.

In both cases, you have in practice to think about what may be the evidence for a particular policy being part of an applicable causal principle. You have also to think about what evidence there may be for this or that being a support factor; and what evidence there may be that those support factors are present in your situation. The nature of the evidence will in general be very different in each case. For example, science may tell you that this policy is part of a causal principle here, whereas the evidence required to know that a known support factor is present may be no more than the evidence of your eyes.
The distinction also matters when you think about what might have gone wrong in Bangladesh. If you spot that the mother was certainly educated and that could have contributed to i.h., but there was not enough, or not the right kind, of food for her education to pay off, then you have case 1.: e_m is in the causal principle, but you did not have b_m'. But if you see that the mother is not in charge of handing out the food, but the mother-in-law is, you have case 2, and e_m is not relevant, so the presence of b_m' does not arise. You might in that case consider the following:

BD: i.h.(i) c= a_1' + a_2' i.h.$_0$(i) + a_3"b_{ml}(i)e_{ml}(i) + a_4'z'(i)

where e_{ml} is educate the mother-in-law. And *a"*, etc. instead of *a'* or *a* signals that you don't know what *a"* stands for and certainly don't know that *a* and *a'* and *a"*, etc. are the same. And the same in the case of the support factors, b_{ml}, for educating mothers-in-law in Bangladesh. They may not be at all the same as for educating mothers in Tamil Nadu, or mothers-in-law in charge of households elsewhere.

If educating the mother-in-law works, you have two equations:

BD: i.h.(i) c= a_1' + a_2' i.h.$_0$(i) + a_3"b_{ml}(i)e_{ml}(i) + a_4'z'(i)
TN: i.h.(i) c= a_1 + a_2i.h.$_0$(i) + a_3 b_m(i)e_m(i) + a_4z(i).

They represent two causal principles with nothing in common except their abstract form. If without doing any more testing you write e_{pw}, meaning *educate the person with the power,* for both e_m and e_{ml}, you get:

TN: i.h.(i) c= a_1 + a_2i.h.$_0$(i) + a_3'"b_{pw}(i)e_{pw}(i) + a_4z(i)
BD: i.h.(i) c= a_1' + a_2' i.h.$_0$(i) + a_3'"b'_{pw}(i)e_{pw}(i) + a_4'(i)z'(i).

These represent two causal principles with one thing (and only one thing) in common, the intervention e_{pw}. Even the support factors for educating the person in power might be different between Tamil Nadu and Bangladesh, hence b_{pw} and b_{pw}' in the two principles.

As II.B shows, this transformation, which appears to amount to no more than inserting mother and mother-in-law each into her equation under another, common, description, is more than just a smart way of creating causal principles that have something in common. As the modest process of reflection described above shows, thinking like that can lead you to the mother-in-law. And in parts of Africa it may lead you to the elder sister.

4. What Makes Effectiveness Predictions True?

Recall our second assumption: To get a good argument from "It works there" to "It will work here", facts about causal principles here and there are needed. Just which facts about causal principles are these? Start with the prediction that the policy will work here. The answer to our question depends on just what is meant by "will work here". What the facts are that will make an effectiveness prediction true depends on exactly what form the prediction takes, that is, on just what you want to predict about the outcome after policy implementation.

Below we present a catalogue for a variety of familiar predictions that you might want to make. For the most part, in this book we focus on the last, minimal, prediction: "The policy will contribute positively here." As you see from the catalogue, this is the weakest of the effectiveness predictions you might want to make, and the factors you need to know about for this conclusion are essential for all the others. For instance, to predict that the overall difference due just to the policy itself is positive, you will need to know about all the positive contributions it makes, and all the negative, and that the positive outweigh the negative. The lessons we propose about warrant for this minimal prediction are then all relevant when it comes to warrant for the stronger predictions. You will need at least the information we propose plus more if you want a stronger conclusion.

- Almost always the most difficult thing to predict, even if you are prepared to admit a reasonable degree of error, is the actual value of the outcome, individual by individual. There are two kinds of facts that are responsible for what the actual value will be for an individual unit *i*:
  – The causal principle for the production of *y* that obtains after implementation.

  - The values for the given individual of all the quantities occurring in that principle.
- Predicting the average outcome across the individuals in the population, though not so demanding, is still incredibly difficult. There are two kinds of facts that are responsible for the average outcome value:
  - The causal principle that obtains after implementation.
  - The averages across the population of all the terms occurring in that principle.
- Often you would be pleased to predict with some confidence that the outcome average will be better if you adopt the policy than if you don't, rather than predicting the actual average that would occur. In many cases this can be easier. But it can sometimes be tricky. The tricky part comes if the causal principle governing the production of the effect changes in the course of policy implementation.

The possibility of changes in the course of policy implementation is an important feature of how the world works, not at all an unusual occurrence. A good many of the causal principles that produce the outcomes of interest in policy deliberation are not basic laws of nature but are derivative. They express what more fundamental principles give rise to in particular structures. When you implement policy you can change these structures and thus change the very principles you were hoping to rely on to predict the outcomes of your interventions.

This is one of the fundamental reasons that the Chicago School in economics urged against government intervention. Robert Lucas, in what has become famous as "the Lucas critique" (1976), made just this claim. He produced models of interacting economic agents that—*if* the models are correct—show that the very fact that the government manipulates a cause, as opposed to the cause taking its "natural" value, will change the underlying structure so that it will no longer give rise to the principle used for policy prediction. His now classic example is of the Phillips curve, which relates inflation to unemployment—a relationship that Lucas argues breaks down when politicians attempt to reduce unemployment by manipulating inflation.

What Lucas claims happens with the Phillips curve can happen anywhere. In our BINP and TINP examples, educating the mother or the mother-in-law may make them feed the children better. But they may also get the idea from other members of the group where they are educated that they might get a job and hand over giving out the food to the eldest child. You have then, perhaps unintentionally, changed the social structure. Educating the mother or the mother-in-law may no longer work because the old causal principles no longer apply given the new social structure. Educating the mother or the mother-in-law will no longer play a causal role in contributing to children's nutrition because it is not in the new causal principle, although educating the eldest child may be.

In cases such as these where the causal principle changes under policy implementation, the average difference will be the difference between the average that would have been produced under the principle that would have obtained were the policy not implemented minus the outcome average under the new principle. So the facts responsible for the average difference are
  - The old principle.
  - The new principle.
  - The averages of all the terms that differ between the two supposing, for the old principle, that the policy is not implemented and, for the new, that it is.
- If the causal principle does not change, the facts that determine the difference between outcome averages are
  - The averages that would obtain were the policy not implemented, for all the terms whose averages change.
  - The averages that would obtain were the policy implemented, for all the terms whose averages change.

  Notice that we mention "All the terms whose averages change". This is important because, as our discussion of the California Class-Size reduction Program illustrates, often in implementing policy you change more causes than just the ones described in the policy, sometimes wittingly, but, if you are inattentive or unlucky, you can easily do so unwittingly. In principles of form CP, the difference in average outcomes with and without the policy will depend on the differences in the averages of the factors represented in *b(i)* and *z(i)*.
- You might be interested in what difference just the policy itself would contribute on average and supposing the causal principles stay the same, rather than what the average difference would be if the policy were implemented versus if it were not. That is, you may not be concerned about the difference in outcome average due to changes in any of the other causal factors that might change during implementation or due to changes in the causal principles themselves. This difference, when evaluated for a situation *S*, is often called the *efficacy* of the policy in *S*. Notice how slight, in terms of our taxonomy, is the significance of efficacy so defined.
  Sometimes you may be interested in the efficacy of the policy in *S* because you think of implementing the same policy again in *S*, keeping the causal principles there fixed, or of implementing it somewhere else where the same principles obtain as obtain in *S*, but this time implementing it in a different way, perhaps in a way that keeps all other causal factors fixed or improves their values. The efficacy of the policy measures just how much the policy itself contributes, averaged over all the individuals in the population under study. How interesting that is depends on how much you can rely on the fixity of the causal principles and factors.

  The factors that determine efficacy are
  - The causal principle that holds, which by hypothesis is supposed to stay the same before and after implementation.
  - The average across the values of the support factors for the policy under that principle.

5. RCTs

We focus on relevance; the well-known ranking schemes focus on trustworthiness, and especially on the trustworthiness of claims that a policy or program worked somewhere, where RCTs and meta-analyses of RCTs are

taken to be the best supporting evidence for such claims. Here we bring together these two projects and show how they relate. The central question for us is: Under what circumstances is an RCT evidentially relevant to an effectiveness prediction? We begin by explaining what an RCT is and why it is thought to be superior to other studies. Then we will investigate just what it is that an RCT can establish. But to be evidentially relevant, an RCT has not just to establish a fact, but offer what any other source of evidence must offer—support for a premise in an argument that leads to a conclusion. So in 6 we will describe how to make RCT results relevant to your effectiveness prediction by showing how to bring them into your argument that the policy will work here, in your setting.

5.1 What is an RCT?

An RCT is a study design based on John Stuart Mill's *method of difference* for making causal inferences (Mill 1843[1850]: bk. III, ch.8). Mill's method-of-difference supposes, as we do here, that effects are produced in accord with causal principles. The causal principles for a given kind of situation or population, *S*, say what the causes of a given effect in *S* are, what each contributes, and how they combine. A method-of-difference study then aims to compare individual units that are the same with respect to all causal factors relevant to the given effect except the one in question, by which they differ. If individuals that are otherwise the same differ in values for the effect, then the factor by which they differ must be among the genuine causes of the effect under the principles governing *S*.

It is difficult to conduct a straightforward method-of-difference study since it is so seldom known what all the factors for a given effect are, bar one. A familiar strategy for coping with this ignorance is to compare, not pairs of individuals, one-by-one, but rather two groups of individuals, one where the putative cause occurs, which is called *the treatment group*, and one where it does not, called *the control group*. Both groups are supposed to be subject to the same causal principle for the effect in question, and the distribution of causal factors other than the one in question between the two groups is (near enough) identical . For principles of form CP, that means that the distributions of y_0, *b*, and *z* are the same. Then if the two differ in the distribution of values for the effect, the putative cause must be a genuine cause for at least some members of the population.

It is naturally difficult to set up such a study since, in general, not all the causal factors relative to an outcome are known, which makes it hard to check that they are distributed equally in the two groups. Moreover differences between the treatment and control are easy to introduce in setting up the two groups (by, for instance, unconscious bias in selecting to which group to assign a individual) or in implementing the treatment (consider, for example, the placebo effect). RCTs are supposed to help with just this problem.

An RCT is a Mill's method-of-difference group study in which individual units, all of which are supposed to be governed by the same causal principle, are randomly assigned to the treatment and control groups. There is also as much masking as possible – those delivering the treatment don't know which group an individual is in, nor those receiving it, those diagnosing whether or to what degree the effect occurs, those doing the statistical analysis, etc. Finally the groups are supposed to be big enough to allow reliable inference from the observed frequencies to the true probabilities.

The random assignment plus masking are supposed to make it likely that the two groups have the same distribution of causal factors. It is controversial how confident these measures should make us that they do this.[5] This issue bears on the trustworthiness of causal claims backed by RCTs. As we noted, trustworthiness is the central topic of many other guides. But we aim to move beyond that; we concentrate on relevance. In order to keep questions of relevance to the fore, let us suppose that these measures do succeed and henceforth focus on the *ideal RCT*: one where no causally relevant differences obtain between the two groups other than the policy and its effects.

5.2 What RCTs Establish

Consider an RCT designed to test what effect, if any, treatment x has in producing outcome y. The standard result from an RCT is the so-called "*treatment effect*", *T,* across the individuals participating in the study (letting *Exp(θ)* represent the expectation of θ and $=_{df}$ mean "is equal by definition"):

$T =_{df} Exp(y(i)|x(i) = X) - Exp(y(i)|x(i) = X')$

where *X* is the value of the treatment in the treatment group and *X'* is its value in the control group.

Of what use is the treatment effect? Here we shall present an argument familiar in the evidence-based policy literature that shows the link between a positive result—a positive treatment effect—in an ideal RCT and the causal conclusion that the result supports. This conclusion, we shall see, is that the policy tested genuinely appears in the causal principle governing the production of the outcome in the experimental situation and that the support factors for it were there for some individuals in the study population; or, in language we use throughout this book, the conclusion that *the policy played a positive causal role* given this principle. It is important for our subsequent discussion that this, although elegantly arrived at, is *all* that it establishes. It is a different and difficult question how that fact can be given the status of evidence for an effectiveness conclusion.

RCT Argument

Following the discussion in I.B.3.1, suppose that y in the study population is determined by a causal principle of form CP:

$$CP: y(i) c= a_1 + a_2y_0(i) + a_3b(i)x(i) + a_4z(i)$$

The formula makes clear that, for any individual i, *b* not only determines in part whether *x* contributes to *y* at all

[5] For a skeptical take on this issue, see (Worrall 2002, 2007).

for *i* but it also, along with a_3, controls how much a given value of *x* will contribute to *y*. From CP and the definition of T it follows that

$$T =_{df} Exp(y(i)|x(i)=X) - Exp(y(i)|x(i)=X')$$

$$= Exp(a_2y_0(i)|x(i) = X) - Exp(a_2\ y_0(i)|x(i) = X')$$

$$+ Exp(a_3b(i)|x(i) = X)X – Exp(a_3b(i)|x(i) = X')X'$$

$$+ Exp(a_4z(i)|x(i) = X) - Exp(a_4z(i)|x(i)) = X').$$

If you are prepared to suppose that the masking and random assignment of individuals to *X* and *X'* assures that for individuals in the study, *x* is probabilistically independent of y_0, *b*, and *z*, as it should be in the ideal, then

$$T = a_3Exp(b(i))(X – X').$$

If T is positive then a_3b is also positive for at least some individuals. So *x* genuinely appears as a cause for *y* in the law CP for the study population. As we say, *x plays a positive causal role* under that principle. If $a_3 = 0$ or *b* = 0 for all individuals then *x* does not appear in CP. So under CP, *x* makes no contribution to *y* outcomes for any individual, it plays no causal role; the outcomes for *y* are produced entirely by the quantities represented in the other three terms in CP.

Notice that the treatment effect averages across the *b* values for different individuals. This has two important consequences. First is the familiar fact that averages conceal what is happening to individuals. A positive treatment effect is perfectly consistent with *x* making substantial negative contributions to *y* for a great many individuals in the population. This is apparently what was happening with the teenage antidepressant that was helpful on average but seemed to make some of those treated with it suicidal (MHRA 2004). Second, although a positive treatment effect shows that *b* must be positive for at least some individuals, the reverse is not the case. A zero average is consistent with exceedingly high values of *b*, both positive and negative, for every individual. So a positive treatment effect shows that *x* can play a positive causal role anywhere the same causal principle obtains, but the lack of treatment effect does not show that *x* cannot play a causal role under that principle.

5.3 Alternatives to RCTs

RCTs are supposed to be the gold standard for evidence in evidence-based policy. Says who? That's the verdict of the usual evidence-ranking schemes recommended for evidence-based policy and used by most policy vetting agencies and policy warehouses. A good example is GRADE, constructed by the Grading of Recommendations Assessment, Development and Evaluation Working Group, which is used by over 50 organizations worldwide (Balshem et al. 2011). You can see their scheme, updated in 2011, in Figure IV.3, along with the definitions of what the various ratings are supposed to mean, in Figure IV.4. Although RCTs are at the top there, better than RCTs, both in GRADE's overall philosophy and in most other ranking schemes, are meta-analyses of RCTs and systematic reviews. Why?

Meta-analyses. Many RCTs have small populations enrolled in the study. This threatens the validity of statistical inference. Suppose there's a difference between the number of positive outcomes in the treatment group versus the control group. That could reflect a genuine difference in outcome probabilities. Or, it could be just an accident, like getting 10 heads in a row in flipping a fair coin. The larger the population, the less likely this kind of statistical accident is. Meta-analyses use statistical techniques to blend together populations from different trials, tending carefully to differences between study designs, to create an imaginary super population in which inference from differences in frequencies of outcomes to differences in probabilities is more secure.

Systematic reviews. These are meant, in the words of the Campbell Collaboration, to ''sum up the best available research on a specific question. This is done by synthesizing the results of several studies.''[6] The studies need not all be RCTs. But they are meant to be "best available", that is, of high quality judged by some explicit, well-grounded criteria. The criteria are given in the evidence-ranking schemes. In our words, the results of the studies included in the review are meant to be highly trustworthy.

Often the reviewers start by looking at dozens, even hundreds, of studies but end up with only a handful that meet the inclusion criteria. This gives rise to some lively debate. Surely it is better, opponents argue, to base a synthetic judgment on all the studies available, taking into account the merits and defects of each. In particular, what if there were two or three high quality studies that pointed one way and a very great many others of varying lesser quality, with a variety of different merits and demerits, that point in the opposite direction? Surely in that case you should not have high confidence in a verdict based on just the few top quality studies. This fits with a standard doctrine about scientific confirmation, based on what is called the *no-miracles argument*.[7] It would be a miracle, so the argument goes, if so many separate, different kinds of defects from different kinds of studies conspired in just the right way to produce similar results. Unless, of course, there were some truth to those results.

Those who advocate considering only the most trustworthy results make two replies. First, "Garbage in, garbage out." Results that are not to be trusted taken as input produce untrustworthy results as output. Second, there's no well-grounded system for "weighing" up evidence of different kinds of different qualities. Too much must be left to judgment, and judgment is not to be trusted.

[6] See http://www.campbellcollaboration.org/what_is_a_systematic_review/index.php.

[7] This argument was originally formulated by Hilary Putnam (1975: 73).

We do not want join this argument. Whatever is the case about including less trustworthy results, it seems hard to quarrel with the idea that a verdict based on a synthesis of trustworthy results will be better than a verdict based on just this or that trustworthy result by itself. So too, it seems hard to quarrel with the idea that a good meta-analysis of studies will be better than the verdict of a few smaller studies. That accounts for why these are at the top of the list. But what about the study designs that appear below RCTs in the rankings? RCTs use Mill's method of difference as their underlying logic. In the GRADE list, "low" and "very low" quality studies do so as well, but without randomization. They are lower in rank because, without randomization, you are supposed to have less assurance that other causally relevant factors have the same distribution in treatment and control groups. Whether this risk is really high or low if you do not randomize, or you do not mask thoroughly at all possible places, depends on exactly what you know about other causal factors, which can sometimes be a lot and sometimes very little. Randomization is often defended by the claim that it is the only way to deal with unknown causal factors. If so, then an ideal RCT can be the superior choice if you are not very secure that you know much about what the significant causal factors are. Supposing that you are in this situation, then ranking good RCT studies above otherwise good studies that do not mask and randomize seems correct—so long as it is remembered as well that what is at stake is trustworthiness, not relevance or cost effectiveness or moral acceptability.

What's surprising, then, is not so much what is immediately above RCTs or immediately below, but what is left out of the usual lists altogether. An ideal RCT can *clinch* the result that the treatment works somewhere. We mean by this that if all the requirements for an ideal study are met, a difference in outcome probability between treatment and control groups deductively implies that the treatment caused the outcome in at least some individuals in the study population. That's what the RCT Argument of 5.2 shows. But there's nothing special about RCTs in this regard. There are many methods where positive outcomes deductively imply causal conclusions, including certain kinds of econometric modeling, process tracing, and causal Bayes-nets methods. Each of these can establish causal conclusions reliably—provided the assumptions backing these study designs are met.

All methods require specific assumptions to be met if the conclusions drawn from them are to be justified. In particular, all methods establishing causal conclusions have assumptions about causality among their assumptions. Hence the slogan, "No causes in; no causes out." All Mill's method-of-difference studies suppose, for example, that every probabilistic dependency has a causal explanation. And they suppose that all causes other than the treatment are distributed in the same way in the treatment and control groups. And they suppose that that means that, if treatment and outcome are probabilistically dependent in the study, there's no explanation left except that the treatment caused the effect in at least some individuals in the study. Other kinds of methods require other assumptions. We describe some of these here, very briefly, to acquaint you with them.

Causal Bayes nets. These use probabilistic dependencies plus any available causal knowledge to infer new causal conclusions. Unlike RCTs, they do not need to suppose the probabilities are from an experimental situation; they can do with data from an ordinary non-experimental population. Not surprisingly then, some of their assumptions are stronger than those required for RCTs. For example, they assume that, once information about a factor's causal predecessors has been taken into consideration, that factor will not be probabilistically dependent on anything except its own effects. Also, it is difficult to get many new results out without the additional assumption that the causally antecedent factors not taken into account are probabilistically independent of each other.[8]

Econometric methods. Econometrics has evolved a number of sophisticated techniques for using probabilities in non-experimental populations to infer functional relations between factors that hold in those populations. But it is well known that not all true functional relations are causal. For instance, if causes are functionally related to their effects, then two effects of the same cause will be functionally related to each other even though neither causes the other. Sometimes, however, the genuine causal relations can be identified. This will be possible if an *instrumental variable* can be found. An instrumental variable is essentially one that affects the cause under test but none of the other possible causes of the putative effect.[9] It is also possible to identify genuine causal relations with other kinds of background causal information, though generally far more background information will be needed.[10]

Process tracing. This method confirms the existence of a causal connection between start and finish by confirming, one-by-one, a series of smaller causal steps in between. For the method to work, the steps in between must either be of a kind that are already well-established or else be ones that can be established on the spot. Sometimes these intermediate steps are not established by direct observation but rather, for instance, by registering side effects that would be produced just in case the effect in question occurred, or by looking for effects of that effect. Process tracing is used regularly in daily life, often to draw negative conclusions. "My baseball couldn't be what broke your window since my baseball never went out of my backyard." And it is a familiar method in both biology and physics. It is also regularly used in *post hoc* policy evaluation. The Carvalho and White (2004) study of social funds, discussed in III.B.2.3 is a case in point.

All of these methods are reliable, so long as their requisite assumptions are met. That

[8] See for instance (Pearl 2009: 146).

[9] For more on instrumental variables, see (Reiss 2005) or (Angrist et al. 1996).

[10] See (Fennell 2007) and (Cartwright 2007).

is, there are arguments just as rigorous as the RCT Argument to show that a causal conclusion follows deductively from positive results. The special advantage of RCTs seems to lie in the fact that few of their assumptions require knowledge about the factors that might be involved or their setting. We have invented the term "self-validating" to label this. In an RCT, you do not need to know a lot of background causal information about this factor or that since the basic assumptions are supposed to be justified by the design of the experiment itself. For instance, you do not have to know the other relevant causal factors to have reason to think they are distributed the same in the treatment and control groups. Randomization, masking, and placebo control are supposed to make this likely.

In many cases, however, there may well be enough information available to support reasonable confidence that assumptions for other methods are met; and in many cases it will be very difficult indeed to conduct a good RCT; and sometimes, as in the Nobel-prize winning work of James Heckman (cf. Heckman and Vytlacil 2007), econometric methods can combine with randomized experiment to give better post hoc evaluations and predictions of policy success. So it is surprising that these other methods are not part of the evidence-based policy canon.

As with all other study designs, these kinds of studies can be done more or less well, and their background assumptions may be more or less trustworthy. They, like the study designs that appear in the usual evidence-ranking schemes, need vetting; and the vetting must be done by experts who know just what to look for. So knowing the existence of these methods is not likely to be of much realistic help to you in your attempts to use good evidence in your policy predictions until the social policy vetting agencies, warehouses, and systematic reviews figure out how to take them into account. In the meantime, much good evidence is being scattered to the winds.

6. Relevance

6.1 What Makes RCT Results Relevant to Effectiveness Predictions?

This depends on the exact form of the effectiveness prediction. But before discussing that, it is important to notice that the treatment effect in an experimental population is not directly relevant to any effectiveness prediction outside the study population; its relevance will always be indirect. We shall for the most part discuss the weakest effectiveness prediction: "The policy will contribute positively if it is implemented here," where this will be determined under the causal principle that will hold here post-implementation. We focus on this, first, because it is the easiest to predict, requiring the fewest further assumptions, and, second, because, as we mentioned, you will need the same information, plus more, for any stronger predictions.

Start with the simplest case, where it can be taken for granted that the study situation and your situation – there and here – are subject to the same causal principle for the production of the targeted outcome *y*. Will *x* make the same average contribution; that is, is the efficacy, which is measured by the treatment effect in the study situation, the same there as here? Certainly if the same principle holds there as here, a_3 will be the same since it is constant. But Exp(*b*(*i*)) is not; it is an average – an average over *x*'s support factors. The average in each situation depends on the distribution of these factors in that situation. Even if the same principles govern the two situations, that provides no reason to suppose that the distributions of support factors are the same. To the contrary, this distribution often depends heavily on local circumstances. So it is unlikely to be the same.

Moreover, the same distribution is not really what you hope for. What you would like is that you have – or can arrange to have – a distribution that favors the good values of *b* – the ones that provide the largest positive contribution from the policy. At the least, you will want to have some values that make *x*'s contribution positive and these should outweigh the effects of those that make *x*'s contribution negative; and if getting negative contributions in some individuals is to be avoided, then you don't want any of these "bad" values of *b* at all.

Laying aside for the moment worries about negative contributions in some individuals, suppose that you want to predict that the policy will contribute positively in your situation. What does it take to make RCT evidence relevant? Or, more broadly, since RCTs are only one way among many to support "It works somewhere", what does it take to make "It works there" (howsoever it is established) relevant to "It will work here"?

6.2 From "It works there" to "It works here"

Suppose "It works somewhere" is trustworthy. When is that evidentially relevant to "It will work here" and under what conditions? From now on, we will take "It will work here" in its weakest sense. Unless we state otherwise, "*x* works in situation *S* with respect to outcome *y*" means "*x* produces positive contributions to *y* for some individuals in *S*". Recall that this allows that *x* may produce negative contributions in other individuals, and may even produce an overall negative average contribution; and even if it produces an overall positive contribution on average, this does not mean that the average will be better than before because of the possible negative effects of other factors that change, either independently or as a result of implementing *x*.

When then will *x* work in S with respect to outcome *y*? That happens exactly when *x* genuinely appears in the causal principle that governs the production of *y* in *S* post implementation and the support factors necessary for *x* to contribute to *y* are present for at least some individuals in S post implementation. In the language we introduced earlier, when *x plays a positive causal role (with respect to y) in S.*

Now we are ready to address the question with which this section began. To know whether "It works somewhere" is evidentially relevant to "It will work here" and under what

conditions, you have to start by asking what kind of argument can go from "It worked there" to "It will work here"? Here is one:

1. *x* works there (ie, *x* genuinely appears in the causal principle that governs the production of *y* there post implementation).

2. Here and there share that causal principle post implementation.

3. The support factors necessary for *x* to contribute under that principle are present for at least some individuals here post implementation.

Conclusion. *x* works here (ie, *x* genuinely appears in the causal principle that governs the production of *y* here post implementation and the support factors necessary for it to contribute to *y* are present for at least some individuals here post implementation).

This argument reflects the fact that "It works there" gives information about the causal principle that obtains there and about the existence of the requisite support factors there. But it gives no information about what the causal principle here is, nor about what support factors, if any, obtain here. You can think, "Surely it is the same here as there." Maybe so. But the issue is, "What do you have warrant for; What degree of confidence are you justified in?", not, "What do you think?" And the answer to the question, "The same how?" matters. Heedful of our remarks in 6.1 about the distribution of support factors, this argument does not suppose the distributions to be the same in both locations. The argument would be valid – the conclusion would still follow from the premises – with that premise substituted for premise 3. But the same distribution is not necessary for the conclusion that it will work for some individuals here; and, as we noted, a "better" distribution here than there would be preferable.

This argument nevertheless demands a lot. It requires the same causal principle to govern the production of the outcome here as there. Recall that a causal principle records the full set of causes that operate, what each contributes, and how they combine. That's what it takes for nature to set the value the outcome will have. But in many domains the causes that operate shift frequently and unpredictably, from locale to locale and from time to time, as economists from John Stuart Mill (1836[1967], 1843[1850]: book VI) to British econometrician David Hendry (Hendry and Mizon 2011) have argued. That's why, said Mill, economics cannot be an inductive science. The principles that held in the past can in no way be relied on to hold in the future, due to the shifting of causes.

But that need not make information about causes there irrelevant to what happens anywhere else. As Mill stressed, many causes have what he called a "stable tendency". They make the same contribution across a variety of different situations; that is, they appear in the same form across a variety of different principles. The forces of physics are a clear example. In different situations – from Galileo dropping balls from the Leaning Tower to airplanes flying at 10,000 feet above the Earth to electrons moving in a battery – gravity always makes a contribution – the same contribution – to the total force exerted on the object, a contribution of size GMm/r^2. It does so no matter what other causes affect the outcome; it *plays the same causal role* in all the different causal principles for all the different situations where masses appear.

Not all causes have stable tendencies. Some seem to operate totally locally. Among those that do have stable tendencies, the range of stability can vary. Perhaps some are universally stable, but most have boundaries. They make the same contribution in a range of situations but not in others, where the breadth of the range can vary dramatically. The trick is to figure out what kinds situations are safely within the range. Within that range you can suppose that the contribution you see there will appear here as well. Ideally this is what science will provide for you. But often the science has not done so, or not done so yet, especially with the kinds of causes at stake in social, economic, and health policy. Then you will have to think about this – seriously – for your special situation and get what advice and help you can.

II.B is all about finding the right kind of causes to be employed in policy design, ones that can make a positive contribution in your setting. A policy that is known to have a stable tendency to contribute positively will fit the bill perfectly. But warranting assumptions about stability of contribution is difficult – it is the meat of serious on-going science. Nevertheless you may have to rely on assumptions like these being true if your policy is to work. For very often the best warrant for the claim that *x* plays a causal role here is that it is already well established that *x* has a stable tendency to produce *y*, stable across a wide variety of kinds of situations, including ones like yours. It produces a positive contribution to *y* in some individuals here because it always – across this range – produces a positive contribution for some individuals, or, even possibly for all individuals. So watch out. This is a difficult assumption to warrant and where the warrant for it is weak, so too is the warrant for any effectiveness predictions it is supposed to support.

We shall, for shorthand, say *x can play a causal role with respect to outcome y* in a situation if *x* genuinely appears in the causal principles only for that situation. For principles of form CP, that means that *x does* (or *will*) *play a causal role in situation S* if it can play a causal role under the principles that govern S and the support factors (designated by *b*) required under those principles take non-zero values for some individuals. Then *x can play the same causal role in situation S' as in S* means that it genuinely appears in the principles for the situation *S* just in case it genuinely appears in those governing situation *S'*, and with the same sign. This is difficult knowledge to come by.

Even if you are reasonably warranted in the assumption that the policy can play a positive causal role in your situation, it is essential to keep in mind that a policy can play both a positive causal role for some individuals and a negative role for others. If this matters, you had better be at pains to learn about both possibilities. And, recall, to predict which

dominates, you will need information about the distribution of values of the support factors for the positive and negative roles.

A final thing to note is that it is generally a whole causal team that has a stable tendency, not an individual cause by itself. Recall, individual causes are generally INUS conditions. Each is usually only part of what it takes to get a contribution. It seems that usually, where there are stable tendencies, the entire team is required to get the stable contribution. Masses, like the sun, cause other masses, like the planets, to experience an attractive force. But what that force is depends not just on the first mass alone (the mass of the sun) – that is only an INUS cause – but on the whole team, which includes the constant of gravity (which is an instance of what we have called a "boost factor"), the mass of the second body (the planet), and the separation between them. You won't get the stable contribution to the force if any member of the team fails to show up for work.

With these considerations in mind, we can construct a different argument for getting from "It works there" to "It will work here", one that does not require the same causal principle to obtain here and there but substitutes for this the assumption that the policy *plays the same causal role* here as there**.**

Effectiveness Argument

1. *x* plays a positive causal role there post implementation.

2. *x* plays the same causal role here as there post implementation.

3. The support factors necessary for *x* to play a positive causal role are present for at least some individuals here post implementation.

Conclusion. *x* works here (ie,, *x* can play a positive causal role here post implementation and the support factors necessary for it to do so are present for at least some individuals here post implementation).

Our task in this section has been to show how to get from results that provide good evidence for "It works there" to the conclusion "It will work here". The task is almost accomplished. All that's needed is to add, as support under the subargument, a sub-subargument, like the RCT Argument of section 5.2, that takes the study results as one of its premises and that concludes with "*x* plays a positive causal role there". If you find the policy vetted by a good policy warehouse like What Works or the Campbell Collaboration, you can take for granted that there is a good argument like this for premise 1. With this addition, the Effectiveness Argument does the job.

This Effectiveness Argument is the one we shall rely on throughout. That's because it is a very special argument. Not only do its premises imply its conclusion, as they should in any good argument. In addition its premises are necessary for the conclusion; the conclusion will not hold without them. Suppose, as we have been taking for granted about the policy under consideration, that it has been shown to work there, say in an RCT study. If the policy does not play the same causal role here as there, it will not work here. Similarly, if the necessary support factors for it are not in place here post implementation, it will not work here. So for the policy to succeed here, premises 2 and 3 must be true. If they are not true, the policy will definitely not work here.

6.3 External Validity

This is a central notion in the RCT orthodoxy, and it does not do the job that it is meant to do. It is meant to deal with the issue that we have raised, whether a policy that has been shown to have worked by a good study can be expected to work in a different context. Every practical person knows, from the high risks and failure often found in rolling out successful pilots, that this is a real problem. So it has to be faced. That's what this book is about.

The orthodoxy approaches this by distinguishing between *internal validity* and *external validity*. A study has *internal validity* when the study provides strong warrant for the study results. The RCT Argument of 5.2 shows that RCTs can provide strong warrant for causal conclusions. There are well established procedures, to do with randomization, masking, and so on, for ensuring that a positive treatment effect – a positive average difference between what happens to those who had the treatment, say a drug or small classes, and those who did not – implies that the treatment played a role in producing the outcome in the study population. We have no quarrel with this nor generally with the notion of internal validity. We think that there are many well conducted RCTs, that many are internally valid, and that the casual conclusions that they show are trustworthy.

In the orthodoxy, a study has *external validity* when the "same treatment" has the "same result" in a specific target as it did in the study. The orthodox advice is that external validity can be expected if the target population is "sufficiently similar" to the study population. For us the key question is how good a job this advice does in getting you from "It worked there" to "It will work here". The answer: you are lucky if it gets you anywhere. First, the advice is vague, surprisingly so given how specific the orthodox guidelines are in assessing RCTs, meta-analyses, and systematic reviews. Second, similarity, if taken seriously, is too demanding; you'd hardly ever be able to export study results if you insisted on similarity. Third, similarity is the wrong idea anyway. Fourth, it is wasteful.

First, *vague*. "Same treatment". Using the same treatment can be fine – so long as you have identified the right description for the treatment. And the right description is the one that plays the same causal role in the target as in the study. Recall our illustration in 3.2. For Bangladesh and Tamil Nadu, that's "educate the person in power", not "educate the mother".

"Same result". What result? Suppose you are interested in getting the "same

treatment effect". This won't happen unless the policy can play the same causal role in the two populations. Let's make that easy by supposing that the two populations are governed by the same causal principle, say a principle of form CP in 3.1. Recall from the RCT Argument that the treatment effect in a population depends on the average of b(i) there. In this case b(i) represents in one fell swoop all the different support factors necessary in the population if the policy is to produce a contribution there. Each separate combination of values of these factors corresponds to a different value of b(i). The treatment effect depends on the average of these values across the study population. Averages depend on the probabilities for the numbers averaged over; so, the treatment effect depends on the probabilities for each different arrangement of values of the support factors, where each different arrangement is represented by a different value, B, for b(i): Prob(b(i) = B).

So, when can you expect the average of b(i) to be the same in the two populations? Represent the probabilities in the two by $Prob_{SP}$ for the study population and $Prob_{TP}$ for the target population. You can expect the average to be the same when $Prob_{SP}(b(i) = B) = Prob_{TP}(b(i) = B)$ for all B's; that is, when all the combinations of values of the support factors have the same probability in the study and target populations. Otherwise it is an accident of the numbers.

So, except for lucky accidents, the treatment effect will be the same in the study and the target only if the policy can play the same causal role in the two populations, the support factors are the same, and the distribution of their values is the same in the two populations. That's a tall order indeed. It is an absurdly tough test to require the same treatment effect as in the study population. If that then is external validity, there is no real chance that a study will have it.

But perhaps "same effect" is to be understood differently. Maybe as "same overall outcome"; or, "same in making a positive contribution in both places". These are different predictions, and, as in section 4, different kinds of facts must be in place for these different predictions to come true. It's these facts you need to know about if you want to predict that you'll get the "same result", not some vague set of similarities.

What about similarity? In what ways are the target and the study to be similar? In all the ways you can think of? Maybe you don't need to bother getting more precise about what similarity means, though, because of our second and third worries.

Second, *similarity is too demanding*. Consider a paper by a team of authors from Chicago, Harvard and Brookings, "What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?" (Ludwig et al. 2008). The paper explicitly addresses the question of where outside the experimental population you are entitled to suppose the experimental results will obtain—for instance, where can you expect higher high school graduation rates for girls in the families that moved. The authors first report "MTO defined its eligible sample as…" We won't write out their long list in this quotation because you can read it in their conclusion:

> Thus MTO data […] are strictly informative only about this population subset – people residing in high-rise public housing in the mid-1990's, who were at least somewhat interested in moving and sufficiently organized to take note of the opportunity and complete an application. The MTO results should only be extrapolated to other populations if the other families, their residential environments, and their motivations for moving are similar to those of the MTO population. (Ludwig et al. 2008: 154-5)

If that's the limit of where MTO results are relevant, maybe it wasn't worth doing the study in the first place.

Third, *similarity is the wrong idea*. Consider the list in the MTO quote. It's a potpourri. The authors seem to have tossed in everything they could think of without system or reason; why for instance did they leave out the geographical location of the cities in the experiment? And anyway, the list does not get at what is necessary. Look at it again. You know that the same treatment effect requires that the policy play the same causal role in the two populations and the support factors have the same distribution in both. Are we really meant to suppose that sharing this long list of factors will ensure that?

Maybe you are to think of the factors in this list just as indicators, where the hope is that sharing the indicators ensures that the two populations share the facts that really matter. But why are these good indicators? If you offer a list of indicators, you need some defence of why the indicators are up to the job. And this will be hard to give without any thought about how MTO plays the role it did in the Chicago study population and of what its support factors might be here.

We are not alone in our demands that you must try explicitly to identify the support factors. Here is what Edward Leamer, an econometrician famous for a classic paper, "Taking the Con out of Econometrics", has to say about it, using slightly different language (his "confounders" are our "support factors"):

> […] the overall treatment effect is not a number but a variable that depends on the confounding effects […]. If little thought has gone into identifying these possible confounders, it seems probable that little thought will be given to the limited applicability of the results in other settings. (Leamer 2010: 35-6)

> […] which is a little like the lawyer who explained that when he was a young man he lost many cases he should have won but as he grew older he won many that he should have lost, so that on the average justice was done. (ibid.)

Fourth, similarity is *wasteful*. The treatment effect averages over arrangements for the

support factors. Some of these arrangements enable the policy to make a big contribution, others only a small contribution. And for others the policy may even be counterproductive. You shouldn't aim for the *same* mix of these arrangements in your population as in the study population. Rather you want a *good* mix – a mix that concentrates on arrangements that allow the policy to do the most for you.

We have been talking mostly about expecting the same treatment effect. That, we said, is a tall order. Why not drop the hope that the policy will produce the same treatment effect and substitute something weaker, like "It will make a positive contribution" or "It will produce an improvement over what would have happened otherwise." Should you be looking for similarities between your population and the study population in these cases? No. You should be looking for what matters to getting the prediction you have in view right. What's wrong with the ideas of external validity and similarity is that they invite you to stop thinking. Why did you decide to try it in Bangladesh? Presumably because you thought that it might work there. Why? Presumably because you had some idea of how and why the policy might work. And that, using our categories, is about causal roles and support factors. Only by thinking in terms of causal roles and support factors can you begin to see what evidence you need if you are going to bet that the policy will work here. You cannot avoid thinking like that. The notions of external validity and similarity are no substitutes.

7. Bottom Line, Put Simply

7.1 Introduction and Apologia

This section aims to give you a simplified version of our theory, so that you can read the rest of the book even if the theory chapter so far has left you confused. As with all simplifications, we cannot pretend that it contains all the detail and rigor of the full version. And it simplifies by, among other things, making what we say rather cruder, so that the skeleton of the analysis is visible without so many qualifications and complexities. This is at a cost, so that even the previously confused reader may wonder at times whether our assertions here can properly be as bald as they are. But that is for clarity.

And in the rest of the book we use the theory to back our discussion. That discussion will inevitably use more or less technical terms, and most of those appear in this section *italicized*. But we have also tried to use simpler language that relies on the theory without always making it rigorously explicit. That too is not without cost. But it may further encourage those who find our theory hard.

7.2 How our Theory Works

This book is about evidence, because it is about evidence-based policy. The point of evidence-based policy is to choose policies that are effective, that will work. And that means will work where and when they are put into effect – what we call *here.* And the general question is, what kind of evidence, and in any particular case, exactly what evidence, will help you with that *effectiveness prediction.* To start to answer that, we need to step back. We need to think about causes and effects and we need to think about evidence.

Will this policy work here, in your setting? Will it cause the result you want where you are? That depends on what the *causal principles* are where you are. Causal principles fix what causes what. Your policy will not produce the outcome you want if there's no causal principle connecting the two. Getting lung cancer is correlated with owning ashtrays. But buying ashtrays will not bring on lung cancer. There's a causal principle, studied in the biomedical sciences, that connects smoking and lung cancer; and there's a causal principle, studied by sociologists and market researchers, that connects smoking and owning ashtrays. But there's no casual principle that connects buying ashtrays and lung cancer.

The causal principles that hold where you are fix two important facts that matter to whether your policy will work there. The first is about *causal roles:* Can the policy play a role in producing the desired outcome in your setting? Smoking can play a causal role in producing lung cancer; owning an ashtray can't. The second is about support factors: What other factors must be in place for it to do so? Maybe smoking only produces lung cancer if one has the right genes. But genes or not, there's no way ashtrays will play a role in producing lung cancer. There's no support factors that can help.

Turn to evidence now. You are looking for evidence to support an effectiveness prediction, a prediction that your policy will work here, where you are. The facts that you need to support this prediction cannot be just any old facts – it is Thursday. They have to be relevant facts. That is what evidence is. It is evidence for something. In this case, for a conclusion that is an *effectiveness prediction*.

And conclusions are the result of *arguments*. That is a technical term. It does not mean, for example, "reason" as in "my argument for going was that I had promised to". It means a chain of reasoning, or more strictly a set of claims – premises – set out in support of a conclusion. The familiar syllogism "*All men are mortal. Socrates is a man. Therefore Socrates is mortal.*" is an argument in which *Socrates is a man* is relevant and therefore, if a fact, evidence by virtue of playing a role in the argument that leads to the conclusion. "It is Thursday" plays no such role.

So to make an *effectiveness prediction* you need to know what facts to verify. They have to be the facts relevant to the prediction – that is what makes them evidence. What is relevant is determined by the structure of the *argument*, the chain of reasoning, that leads to the prediction. Because that *argument* is about what causes the outcome in your setting, its structure will reflect facts about the *causal principles* that hold there, in particular, facts about *causal roles* and *support factors.*

Having promised simplicity, we may in the last paragraph have delivered only brevity without clarity. So in the rest of this section, we use some of the analysis from the earlier parts of

the chapter to put (back) some of the flesh on the bones of the discussion. But before that, two further points.

We say that "you need to know what facts to verify". Our book tries to help with that question. But it does not tell you how to verify the facts you need once you have identified them. The distinction is important. If you want to drive off by pressing the accelerator, you rely on a causal principle in which the presence of fuel plays a causal role. We help you to realize that, therefore, *There is fuel* is relevant, and that you need evidence for it. We do not help with how you find out whether there is fuel, how you get that evidence. In this toy example, finding out is trivial – look at the fuel gauge or dip the tank. Often it is not trivial. At worst, it may not even be clear how you would verify some facts. But anyway, we do not deal with that. And the main reason is that there are a very large number of ways of establishing facts, and hence evidence, from common observation to elaborate statistical research.

Second, we spend time here using our theory of evidence to discuss what is the relevance of the fact that a good RCT has shown that the policy worked there. We show that, although this may indeed be a fact, whether it is relevant, and hence evidence, depends on what *argument* it contributes to. And it turns out that its relevance will often be slight.

7.3 Causal Principles

"Will this policy work here?" The question focuses attention on the policy. But to answer it, your focus must be directed elsewhere. Where? The answer to that is supplied by considering some general facts about how causes work to produce their effects. Causes do not produce their effects willy-nilly but for a reason. They produce effects in some systematic way, in accord with some causal principles. A causal principle for a situation lays out all the factors that operate to bring about the outcome in question in that situation and shows how these combine to produce it. We stress three important facts about causal principles that matter to getting the right prediction about whether a proposed policy will work for you:

1. The causal principles that underwrite policy prediction are not universal.

2. Few causes work on their own; causal factors work together in teams.

3. There are generally a number of distinct teams at work in any situation, each making its own contribution to the effect.

1. *Causal principles are not universal.* They differ from place to place and from time to time. That means that it is not enough for you to know that the policy worked somewhere or even that it has worked at some time here. "It worked there"; it played a positive causal role there. So it was one of the factors from a causal principle that holds there. To predict that it will work here, you need to know that it is one of the factors from a causal principle that holds here. That is what ensures that it can play a positive causal role for you. You will read more about finding factors that can do so in II.B.

2. *Causes work in teams.* What gets highlighted as the *cause* – where for you that means your policy – is rarely enough to produce a contribution to the effect on its own. It needs team support. If any of the essential team members is absent, the policy won't make any contribution at all. It is like trying to make pancakes with no baking powder. So even if you know, maybe from a good RCT, that the policy worked there and that the same causal principles hold here as there, that is not enough to conclude that it will work here. That only shows that it *can* play a causal role here. To know that it *will* play a positive causal role here, you also need to know that you will have the requisite support factors here when you need them. That's what II.A is about.

3. *Distinct teams produce distinct contributions.* The overall effect achieved is usually made up of separate *contributions* from a number of different teams of causes, some of which can pull in different directions. The magnet, in team with the iron in the pin, contributes an upward force on the pin; the pull of the earth, in team with the pin's mass, contributes a downward force. The overall force is a combination of the two. Social causes are just the same. Some improve the outcome in view; others contribute negatively to it. This makes predicting the actual outcome difficult. To predict that, you must take account of all the factors at work and of how they combine; that is, you need to know the full causal principle.

Our concern is with less ambitious predictions than "What will the actual outcome be?". We focus on "Will this policy make a positive contribution?" Will it make things better for some individuals than they otherwise would be? In our terminology, "Will the policy *play a positive causal role* here?"

We talk more about the effects of other teams, both ones that do and ones that do not include the policy, in II.A.

7.4 Evidence, Argument, and Warrant

No-one can doubt that basing your predictions about policy effectiveness on evidence is a good idea. But what counts as evidence – good evidence – that a policy worked there or that it will work here? Our answer is theory based. It is grounded in a systematic account of evidence, knowledge, and warrant.

Some claims are self-evident or already well established. They do not need to be backed up by anything further for you to be justified in taking them to be true. Policy effectiveness predictions are not like that. They need support. What does the support look like? To justify having a high degree of confidence in a questionable claim, a claim that is not self-evident, you need to produce some further claims that, taken together, ensure that the questionable claim is likely to be true. That is, you need a good argument. An *argument*, recall, is a set of claims – premises – offered in support of a conclusion. A *good argument* is one in which the premises themselves are all well warranted – trustworthy – and together imply the conclusion, or at least make it highly likely. (In technical language, the

first means that the argument is *sound*; the second, that it is *valid*.)

What holds in general holds for the particular case of policy effectiveness conclusions. There is nothing special about them in this respect. Suppose, then, you have trustworthy information that a policy works somewhere, or in a number of somewheres. So you can take for granted "It works there". What does it take for this information to count as evidence for "It will work here"? It takes a good argument, an argument in which the conclusion—"It will work here"—genuinely follows from the premises, including the premise "It works there", and the premises themselves are trustworthy. It is important to keep in mind that the conclusion of an argument can be no more trustworthy than any of its premises: dicey premises yield dicey conclusions. So identifying what all the premises are matters.

What then about evidence? Suppose you had access to a gigantic encyclopedia that reported every true fact there is. Which facts should get labeled "Evidence for my conclusion" and which not? Evidence is supposed to help justify taking your conclusion to be true, and what it takes to do that is a good argument. So facts can't be just labeled "Evidence for this conclusion" one way or another. It takes a good argument to connect the two. Any fact that gets so labeled will have to be one among many, possibly very many, premises that are each themselves well warranted and that together make the conclusion probable. This makes evidence highly conditional. No matter how trustworthy a fact from the encyclopedia or a result from a scientific study is, it provides no justification at all without the other premises to connect it to the conclusion.

7.5 Arguments for Effectiveness

For evidence-based policy you are urged to use only policies that have been shown to work somewhere. If you are lucky you will be able to find lists of such policies in a warehouse that vets studies and ensures that the claim "It works somewhere" is trustworthy. What does it take to turn that into evidence that it will work here? A good argument that takes "It works somewhere" as a premise and concludes with "It will work here".

There are a number of different arguments that can do the job, with more and less demanding premises. The Effectiveness Argument from 6.2 is the argument of choice for evidence-based policy. It has the weakest premises and, what really matters, its premises are essential to the truth of the conclusion. If either premise 2 or premise 3 of this argument is false, the policy that worked there will not work here. We express this argument in a very simple, straightforward form here.

The basic argument that links "It worked there" with effectiveness predictions looks like this:

Effectiveness Argument

1. The policy worked there.

2. The policy can play the *same causal role* here as there post implementation.

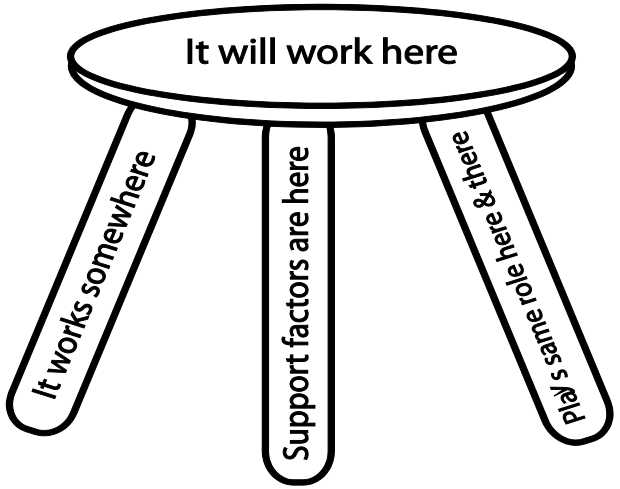
3. The *support factors* necessary for the policy to play a positive causal role here are in place for at least some individuals here post implementation.

Conclusion. The policy will work here.

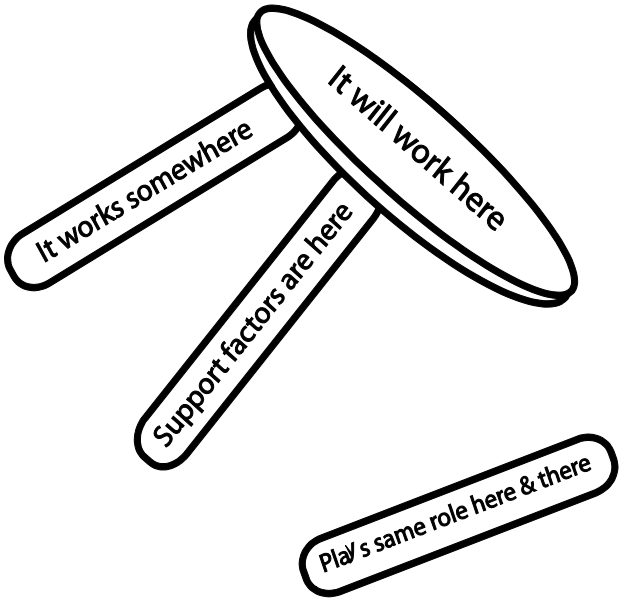
By support factors, we mean the other members of the team of causes that will be required for the policy to make a contribution. Of course it is going to be difficult to warrant a claim like premise 3 without any commitment about what the support factors are. So you will probably have to have a subargument that defends your claims about what these factors are in your situation.

4. So there are the two further pieces of information besides "It works there" that are needed to justify taking "It will work here" as true. These are our chief focus in this book. We explain about them in detail, and with examples; the first, which concerns causal roles, in II.B and the second, which concerns support factors, in II.A. In III we provide advice for how to go about identifying the information these premises demand.

Our central claim in this book is made graphic in Figures I.5 and I.6. If you start, as you are urged to do, with evidence that the policy you are considering worked somewhere, the prediction that the same policy will work here, in your situation, is like the seat of a three-legged stool. It doesn't matter how sturdy one leg is, if either of the others fails, the stool collapses and you are dumped flat on the ground.



[Figure I.5]



[Figure I.6]

7.6 RCTs and Effectiveness Predictions

Randomized controlled trials (RCTs) are supposed to be the gold standard in evidence-based policy. So we want now to apply our theory to the claim that a good RCT is strong evidence for an effectiveness claim.

A good RCT with a positive outcome does indeed show that the policy worked somewhere; that it made a positive contribution for some individuals there, then, in the study population under the study conditions. In our language, "It played a positive causal role there." That tells you something about the causal principles that obtained there. It guarantees that the policy appears in a team that contributes positively according to those principles and that all the other members of the team – the *support factors* – were present there for at least some individuals in the study population.

That information is not much use to you if, as we discuss in II.B, the causal principles for your situation are very different from those of the study situation. So, to use this RCT result to back up a prediction that the policy will work here in your situation – that it will play a positive causal role here – you need warrant for assuming that the principles that hold here are sufficiently like those that hold there. They do not need to be exactly the same. You may have a different mix of causal factors at work.

But they definitely must be the same in this respect: The policy appears in both principles and in a team whose contribution is positive. Whatever is true about what other causes you have here and whatever they have there, if the policy is to make a positive contribution here, it must appear in the principles here just as it did in the principles there.

If the principles are the same in this respect, we say that the policy *could* play a positive causal role here. We say "could" for a good reason. Whether it will or not depends on whether you will have here all the other members of the team needed to support it according to the causal principles that obtain here, and at the right times. If you know that's the case, you can predict with high confidence that the policy will play a positive casual role here just as it did in the RCT.

So a positive outcome in an RCT shows that the putative cause *did play* a causal role in the study situation. It did so because it *could* play a positive role there – it appears in a team that makes a positive contribution under the causal principles that hold there; and because, for some individuals in that situation, the support factors for it were in place. That's what we mean by "*It worked there*" – there, in the study population. Hence, "*It works somewhere*"

There is also the case where it *could play* a causal role but does not because not all the factors in the support team are present. We want to make this distinction between *did play* and *could play* to preserve the insight that an intervention may have the potential to play a causal role in producing an outcome, even if it does not because not all the support factors are present. So small class sizes *could* play a causal role in producing better reading scores in California, even if that policy did not play a causal role because there were not enough good teachers to take the larger number of classes. *Could* then refers to the fact that smaller class sizes did at least figure as part of the causal principles, that is, as part of a story about relevant factors and how they combine to produce the effect. Unlike eating tomatoes twice a day, which does not figure as a part of a causal principle for reading scores, so forget support factors.

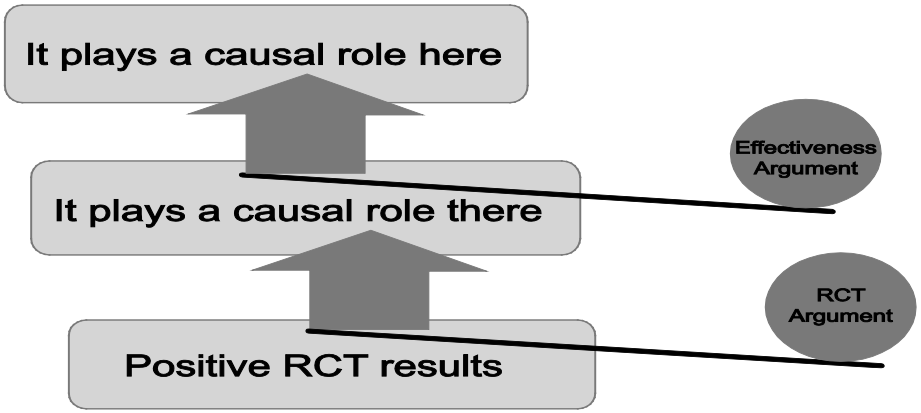
To turn RCT results into evidence it takes a good argument – an argument with trustworthy premises from which the conclusion genuinely follows. We have proposed above in 7.5 an argument – the Effectiveness Argument – that gets you to the conclusion you want, that the policy will work here. But you can see that RCT results do not figure anywhere this argument. Where do they enter? They figure as a premise in a subargument – the RCT Argument – for a premise in the Effectiveness Argument, as in Figure I.7.

The RCT Argument (which we described in 5.2) starts with the claim that the results were positive in the study and concludes that the policy played a positive causal role there – "It worked there". The other premises in the RCT Argument describe the design of the study and provide the connection between causes and probabilities that allows a causal

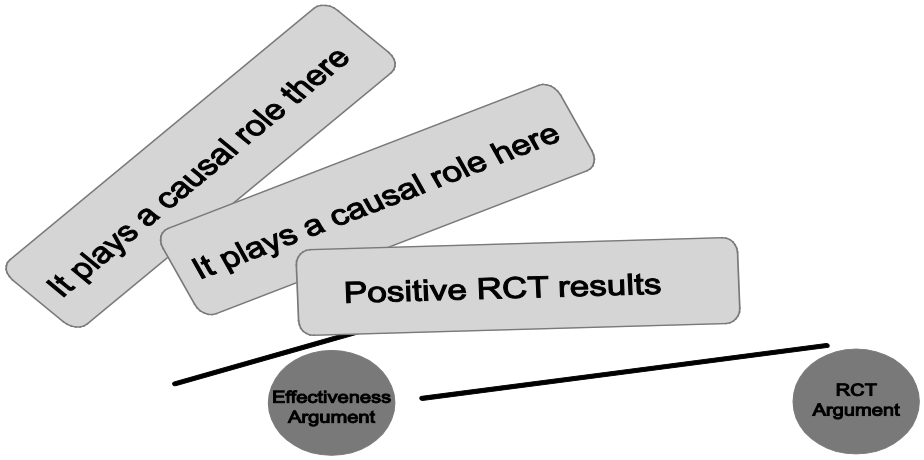
conclusion to be derived from a difference in the average values of the outcome in the treatment and control groups.

We do not focus on the RCT Argument since there are a number of good warehouses, like What Works and the Campbell Collaboration, that take on the job of policing RCT Arguments for you. Though they may not put it this way, what they do is to check to make sure, for specific studies, that the other premises in the RCT Argument are trustworthy, so that a positive result in the study does indeed support the claim that the policy played a positive causal role there. What we are concerned with are all the things that have to be established once the RCT Argument is in place – the other premises in the Effectiveness Argument. This is what our lessons are about.

The important lesson about RCTs is that the relevance of RCT results is highly conditional, depending on both the Effectiveness Argument and the RCT Argument.



[Figure I.7]



[Figure I.8]

As in Figure I.7, a positive result in an RCT is leveraged into evidence that "It works there", there in the RCT situation, by the RCT Argument. Then "It works there" is leveraged into evidence for "It works here" by the Effectiveness Argument. If either of these arguments fails, as in Figure I.8, the lever drops. The evidential relevance disappears with a thud. Worse, you will end up with a policy that does not work for you.

References

Angrist, J., Imbens, G., and Rubin, D. (1996). 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association*, 91: 444-455.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Fennell, D. (2007). 'Why Functional Form Maters: Revealing the Structure in Structural Models in Econometrics', *Philosophy of Science (Supplement)*, 74: 1033-1045.

Mackie, J. L. (1965). 'Causes and Conditions', *American Philosophical Quarterly*, 2: 245-64.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. New York: Oxford University Press.

Putnam, H. (1975). *Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Reiss, J. (2005). 'Causal Instrumental Variables and Interventions', *Philosophy of Science (Supplement)*, 72: 964-976.

Worrall, J. (2002). 'What Evidence in Evidence-Based Medicine?', *Philosophy of Science*, 69: S316-30.

———. (2007). 'Why There's No Cause to Randomize', *British Journal for the Philosophy of Science*, 58: 451-88.

4.2 WILL THIS POLICY WORK FOR YOU? PREDICTING EFFECTIVENESS BETTER: HOW PHILOSOPHY HELPS

(Philosophy of Science Association Presidential Address, 2010 and Philosophy of Science 79 (5), 2012)

1. A Focus on Development Economics

The World Bank estimates that in developing countries 178 million children under five are stunted in growth and 55 million are underweight for their height (World Bank 1995). Malnutrition leaves children vulnerable to severe illness and death and has long-term consequences for the health of survivors. The Bank has funded a wide range of nutritional interventions in developing countries, in Latin America, the Caribbean, Africa and East and South Asia. This included the Bangladesh Integrated Nutrition Project (BINP), modeled on its acclaimed predecessor, the Indian Tamil Nadu Integrated Project (TINP). What was integrated? Feeding, health measures and, centrally, education of pregnant mothers about how better to nourish their children.

TINP covered the rural areas of districts with the worst nutritional status, about half the Tamil Nadu state, with a rural population of about 9 million. Malnutrition fell at a significant rate. The World Bank Independent Evaluation Group concluded that half to three fourths of the decline in TINP areas was due to TINP and other nutrition programs in those areas.

The Bangladesh Project was modeled on TINP. But Bangladesh's project had little success. A Save the Children UK assessment concludes that program areas and non-program areas still had the same prevalence of malnutrition after 6 years and this despite the fact that the targeted health educational lessons sank in to some extent: Carers in the BINP areas had on the whole greater knowledge about caring practices than those in non-BINP areas. Why then did the project fail in Bangladesh?

Before that we had better ask: "Why should it have been expected to succeed?" The extrapolation to Bangladesh from uncontroversial success in India was not warranted, I shall argue, because it was based on simple induction; and simple induction is no better method in social science than in natural science and no better in policy science than in pure science. Moreover, we can do better, and often with knowledge already at hand.

My talk will concentrate on development economics and on a vigorous take-over movement fast gaining influence there, a new methodology to improve development outcomes: randomized controlled trials. As a Public Radio International interview reports: "A team of economists at MIT says it's time for a new approach – one that makes prescriptions for poverty as scientifically-based as prescriptions for disease."[1] MIT's Esther Duflo is one of the leaders of this movement. She tells us:

> The past few years have seen a veritable explosion of randomized experiments in development economics. (Duflo and Banerjee 2009, 152)
>
> Creating a culture in which rigorous randomized evaluations are promoted,

[1] http://www.pri.org/theworld/?q=node/10887

> encouraged, and financed has the potential to revolutionize social policy during the 21st century . . . (Duflo, quoted by *The Lancet* 2004, 731)

Witness also the recent *Journal of Economic Perspectives* symposium on a paper commending RCTs by my LSE colleague Steve Pischke and another MIT economist, Joshua Angrist. As one exemplar of good research design they cite:

> . . . in a pioneering effort to improve child welfare, the Progresa programme in Mexico offered cash transfers to randomly selected mothers, contingent on participation in prenatal care, nutritional monitoring of children, and the children's regular school attendance . . . . Progresa is why now thirty countries worldwide have conditional cash transfer programmes. (Angrist and Pischke 2010, 4)

That's serious extrapolation!

And, to see why I am concerned: Even since I wrote this in draft I have learned that the father of Progresa, Santiago Levy says that many of the places that want them are places where they will obviously fail. In some of these countries success would require people to go to clinics that don't exist (Deaton 2010, 449).

Here's another, from the Jamil Poverty Action Lab (J-PAL), which Duflo and other MIT economists work with: the Deworm the World Movement. J-PAL website reports that "Research by J-PAL associates . . . Kremer and . . . Miguel has shown that school-based deworming is one of the most cost-effective methods of improving school participation." The Kremer and Miguel study looked at 75 primary schools in Busia Kenya. Busia, the J-PAL website explains, "is a poor and densely-settled farming region in western Kenya adjacent to Lake Victoria. [It has] some of the country's highest [intestinal worms] infection rates, in part due to the area's proximity to Lake Victoria Kenya." The website goes on:

The evidence from [the Kremer and Miguel] study has helped inform the debate and has contributed to the scale-up of school-based deworming across 26 countries where over 7 million children have been dewormed since 2009.[2]

I focus on development and on RCTs. But the problem of using evidence of efficacy from good studies and pilots to predict whether a policy will be effective if implemented is a general one. And it is a mega problem. It affects us all. This mega problem, like a good many other problems involving the practice and use of science, is one philosophers of science can contribute to. We are in a position to step in and help, and we should. If we don't step forward to act to improve the decisions that influence all our lives, what is philosophy good for? So let's look at some philosophy that can help. I start with a familiar philosophical concern.

[2] http://www.povertyactionlab.org/scale-ups/school-based-deworming

2. Let's get straight what we are talking about

RCTs, proponents argue, are the "gold standard" for warranting causal claims. But there's startlingly little attention to what these claims claim. In particular there's widespread conflation of 3 distinct kinds of causal claims. RCTs are especially good only for the first:

1. It works somewhere.
2. It works in general.
3. It will work for us.

Here's a typical example from a paper by Duflo and Kremer. Already in line 5, in one single sentence, all three kinds of claims are mixed together without note:

> The benefits of knowing which programs work . . . extend far beyond any program or agency, and credible impact evaluations . . . can offer reliable guidance to international organizations, governments, donors, and . . . NGO's beyond national borders. (Duflo and Kremer 2005, 205)

I take it from the language and use that they mean:

Which programs work = It works in general
Impact evaluation = It works somewhere
Reliable guidance = It will work for us

I focus on these three kinds of causal claims because I endorse evidence-based policy and I want to improve policy outcomes by the use of evidence. The first –it works somewhere – is where we are encouraged by evidence-based policy guidelines to start. These are the kinds of claims that our best scientific study designs can clinch. The third is where we want to end up – the proposed program will produce the desired outcome *in the target situation* and *as it will be implemented there*. The middle –g "general" causal claims – is the central route by which "It works somewhere" can make for evidence that it will work for us. But the road from "It works somewhere" to "It will work for us" is often long and tortuous. There are four essential materials for building a passage across:

> 1. *Roman laws*. I call them this on account of Luke 2.1: "And it came to pass in those days, that there went out a decree from Caesar Augustus, that all the world should be taxed." The laws involved need not be really universal. But they must be wide enough to cover both the evidence and the prediction the evidence is evidence for.
>
> 2. *The right support team*. We need all those factors without which the policy variable cannot act.
>
> 3. *Straight sturdy ladders*. So you can climb up and down across levels of abstraction without mishap.
>
> 4. *Unbroken bridges*. By which the influence of the cause can travel to the effect.

You must have all four; if any one is missing, you can't get there from here.

3. What's an RCT and what's it good for?

I would hope to stay away from formulae in an address like this but we do need some technical results to get started.

An ideal RCT for cause X and outcome Y randomly assigns individual participants in the study, {u}, into two groups where X = x universally in the treatment group and X = x' ≠ x universally in the control group. No relevant differences are to obtain in the two groups other than X and its downstream effects. The standard result measures the average "treatment effect" across the units in the study: So T average is the average of Y in the treatment group minus its average in the control group. Of what interest is this strange statistic about randomized units in a study group?

Supposing that Y values for the units in the study are determined by a causal principle that governs the study population, the RCT can reveal something about the role of X in this principle. Without significant loss of generality we can assume that the principles governing Y look like this:[3]

L: $Y(u) c= \alpha(u) + \beta(u)X(u) + W(u)$

where W represents the net contribution of causes that act additively in addition to X and where X may not play a role in the equation at all if $\beta = 0$. So doing a little algebra (and letting $\langle\Phi\rangle$ represent the expectation of Φ),

$\langle T\rangle =_{df} \langle Y(u)/X(u)=x\rangle - \langle Y(u)/X(u)=x'\rangle$

$= \langle\alpha(u)/X(u) = x\rangle - \langle\alpha(u)/X(u) = x'\rangle + \langle\beta(u)/X(u) = x\rangle x - \langle\beta/X(u) = x'\rangle x' + \langle W(u)/X(u) = x\rangle - \langle W(u)/X(u) = x'\rangle.$

Suppose, as is hoped, that the random assignment of u's to x and x' implies that for u's in the study, X is probabilistically independent of α, β, and W. Then,

$T = \langle\beta(u)\rangle (x - x').$

Recall L: $Y(u) c= \alpha(u) + \beta(u)X(u) + W(u)$

So $T \neq 0 \rightarrow$ X is a contributing cause for Y in L.

You don't really need to follow the details here; just note the bottom line: If the standard assumptions for an ideal RCT are met, the average treatment effect is the difference in X between treatment and control times beta average. So if the average treatment effect is positive then β is too, in which case X genuinely appears as a cause for Y in law L. This, however, provides no evidence that X will produce a positive difference in the target unless the target and the study share L.[4] L must be general to at least that extent. But the stretch of L is in no way addressed in the RCT and for the most part generality cannot be taken for granted. That's because the kinds of causal principles relevant for policy effectiveness are both *local* and *fragile*.

4. Roman Laws are not all that easy to come by

The causal laws we rely on for reliable predictions in real policy, real technology, and real experimental settings are *local*. They are local because they depend on the mechanism or the social organization, what I have called the "socioeconomic machine",[5] that gives rise to them. Economists know about this kind of locality. The Chicago School notoriously used it as an argument against government intervention: The causal principles that governments have to hand to predict the effects of their interventions are not universal. They arise from an underlying arrangement of individual preferences, habits, and technology and are tied to these arrangements. Worse, according to the Chicago School, these principles are *fragile*. When governments try to manipulate the causes in them to bring about the effects expected, they are likely to alter the underlying arrangements responsible for those principles in the first place, so the principles no longer obtain.[6]

Or, British econometrician Sir David Hendry urges the use of simple "quick catch-up" models for forecasting rather than more realistic causal models because the world Hendry lives in is so fluid that yesterday's accurate causal model will not be true today.[7] J.S. Mill too. Economics cannot be an inductive science, he argued, because underlying arrangements are too shaky; there's little reason to expect that a principle observed to hold somewhere sometime will hold elsewhere or later because there's no guarantee the underlying arrangement of basic causes will be the same.[8]

Because so many of the causal principles we employ are tied to causal structures that underpin them, you can't just take a causal principle that applies here, no matter how sure you are of it, and suppose it will apply there. After all, common causal structures are not all that typical, even in the limited and highly controlled world of structures we engineer. Consider for instance these three toasters I found on sale in Oxford: the Cuisinart Classic 4-slice at 41.46 GBP, the Krups expert black and stainless steel at 44.99 GBP, and the Dualit 3-slice stainless steel at 158.03 GBP. Even these three toasters – man-made and for the same job – do not have the same structure inside. (Or at least we hope not given the big price differential!)

Perhaps you think – as many other economists and medical RCT advocates seem to – that the different populations you study, here and there, are more likely to share causal structure than are toasters. That's fine. But to be licensed in that assumption in any given case you better be able to produce good evidence for it.

Simple induction is no more warranted here than anywhere else. It requires stable principles, and stable principles require stable substructures to support them. Without at least enough theory to understand the conditions for stability, induction is entirely hit or miss. This I take it is a key point of Princeton economist Angus Deaton's British Academy Keynes lecture in economics. He says of RCTs that they are,

> unlikely to recover quantities that are useful for policy or understanding. Following Cartwright . . . I argue that

[3] The important lessons follow equally for more complicated functional forms.

[4] Or at least share the important feature of L that X genuinely appears in it.

[5] See (Cartwright 1989).

[6] See (Lucas 1976).

[7] See (Hendry and Mizon 2011).

[8] See (Mill 1836/1967, 1843/1850: book VI).

> evidence from randomized controlled trials has no special priority . . . the analysis of projects needs to be refocused towards the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work . . . thirty years of project evaluation in sociology, education and criminology was largely unsuccessful because it focused on *whether* projects work instead of on *why* they work.[9]

Moving on. Let's suppose though that:

*there are causal principles that enable X to produce Y in the study,
*these are shared in the target, and
*contrary to expectations from the Chicago School of economics, these principles will be unaffected if the proposed policy is implemented in the target.

There are still three central problems for the prediction that the policy will work in the new setting. The next problem concerns the *support team* necessary if X is to produce a contribution to Y.

5. Support Teams

Return to the abstract form L for the causal law that, for purposes of argument, we are now taking to be shared between study and target situations:

L: $Y(u) c= \alpha(u) + \beta(u)X(u) + W(u)$.

The RCT tells about β. It is tempting to think of β as a constant or as an undecomposable random variable. But it isn't. And this despite the fact that you can find it treated thus in sundry works in our field (maybe not from A to Z but at least from Cartwright to Woodward). The difference depends on the kinds of factors that the variables represent. When I write β as a constant or a random variable I assume that "X" represent a *full*, not a *partial* cause. But most policy variables represent only partial causes – INUS causes, extending J.L .Mackie's sense[10] to multivalued variables:

> X is an INUS contributor to Y: X is an *i*nsufficient but *n*onredundant part of a complex of factors that are *u*nnecessary but together *s*ufficient to produce a contribution to Y.[11]

What matters here is that policy variables are rarely sufficient to produce a contribution – they need an appropriate support team if they are to act at all. The support factors are represented by β.[12] And the values of these factors can be expected to vary across the units just as the values of X and W vary.

[9] Read at the Academy 9 October 2008 and published, in a revised form, as (Deaton 2009).

[10] See (Mackie 1965).

[11] "Contributions" are, at least as I make sense of them, defined relative to a metaphysics of capacities, other contributions, and laws of composition. In a law of form L, each separate additive term on the right hand side represents a contribution. See (Cartwright 2009).

[12] In this case we are supposing that the size of the contribution of X to Y is fixed once the values of the "helping factors" are set. But this contribution could still vary arbitrarily from unit to unit. It would be more usual though to suppose that a full set of helping factors would at least fix the probability for a contribution of a given size.

This is well-known in philosophy and in social science. Nevertheless the consequences are frequently ignored. Consider for example the usual advice in the evidence-based policy literature about how to grade policy proposals on the basis of evidence. The US Department of Education explains that what you need are successful RCTs in two or more typical school settings, including "school settings similar to yours" (USDE 2003, 10). And SIGNs, used to help set best practice for the UK National Health Service, provides an A grade to a policy if it is supported by "At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population . . ." (SIGN 2011, 51) This advice is vague, surprisingly so given how specific the guidelines are in assessing RCTs, meta-analyses and systematic reviews. Moreover, if properly spelled out, it is hard to follow. Worst, it is generally bad advice.

Start with hard to follow and consider a paper by a team of authors from Chicago, Harvard, and Brookings, "What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?" The paper explicitly addresses the question of where outside the experimental population we are entitled to suppose the experimental results will obtain. The authors first report "MTO defined its eligible sample as . . ." I won't read their long list because I am about to cite it in their conclusion:

> Thus MTO data . . . are strictly informative only about this population subset – people residing in high-rise public housing in the mid-1990's, who were at least somewhat interested in moving and sufficiently organized to take note of the opportunity and complete an application. The MTO results should only be extrapolated to other populations if the other families, their residential environments, and their motivations for moving are similar to those of the MTO population. (Ludwig et al. 2008, 154-155)

The list is a potpourri. It seems as if they have tossed in everything they can think of that might matter without any systematic grounds; why for instance did they leave out the geographical location of the cities in the experiment? And anyway, the list gets at what's necessary indirectly.

Look again at β in principle L and in the treatment effect:

L: $Y(u) c= \alpha(u) + \beta(u)X(u) + W(u)$

$<T> = <\beta(u)> (x - x')$.

β represents in one fell swoop all the different supporting factors necessary if X is to contribute to Y . Each separate combination of values of these factors corresponds to a different value of β. The average treatment effect depends on the average of these values across the study population. That means we suppose that each different arrangement of values of the supporting factors represented by a different value, b, of β appears in that population with a specific probability: $Prob_{SP}(\beta = b)$.

So, supposing L obtains in both the study and target populations, when can we expect $<\beta(u)>$ to be the same? Exactly when $Prob_{SP}(\beta$

= b) = $Prob_{TP}(\beta = b)$ for all b's; ie, when all the combinations of values of the supporting factors have the same probability in the study and target populations. Otherwise it is an accident of the numbers. I expect that the distributions in the study population are rarely duplicated in other populations.

Independent of that, the list in the MTO article does not seem to be a list of supporting factors. Perhaps the hope is that the list includes sufficient "indicator" factors to ensure that populations that share these indicators will have the same probability distributions over β. Maybe sometimes this is the best we can do. But if we resort to it, we need some defense of why the indicators might be up to the job. And this will be hard to provide without explicit discussion of what the supporting factors might be.

Suppose though we solve the problems of identifying these factors. Still advice like that of the Department of Education is wasteful. The treatment effect averages over arrangements for the supporting factors. Some of these arrangements enable X to make a big contribution, others only a small contribution, and for others X may even be counterproductive. We shouldn't aim for the *same* mix of these arrangements as in the study population but rather for a *good* mix – a mix that concentrates on arrangements that allow X to do the most for us.

I am not alone in this view. In 1983 Edward Leamer wrote a classic paper, "Taking the Con out of Econometrics". The symposium discussing the Angrist and Pischke paper was called "Con out of Economics". Leamer's contribution to that symposium makes the same point about supporting factors I have long argued. Here are Leamer's words:

> With interactive confounders [NC's "supporting factors"] explicitly included, the overall treatment effect [our <β>] is not a number but a variable that depends on the confounding effects . . . . If little thought has gone into identifying these possible confounders, it seems probable that little thought will be given to the limited applicability of the results in other settings.
>
> . . . which is a little like the lawyer who explained that when he was a young man he lost many cases he should have won but as he grew older he won many that he should have lost, so that on the average justice was done. (Leamer 2010, 34-35)

For a final example of sensitivity to supporting factors, return to the integrated nutrition program. The need for getting the requisite supporting factors into place was not ignored in either Tamil Nadu or in Bangladesh. One of the central ideas of the nutrition program was that better nutrition can be secured with meager resources, but to do so, mothers need to know what makes for good nutrition. On the other hand nobody expects that education is enough by itself. You can't feed children better if you can't feed them at all. So the educational program for mothers was coupled with a supplemental feeding program. Nevertheless the results were disappointing. To see what is supposed to have gone wrong, despite the presence of a good support team, turn to my 3rd problem: *ladders*.

6. Ladders

I am a pluralist and a particularist, inclined to suspect that everything is different. Economists are often more homogenizing (though not Hendry and Mill!). They believe that they can base their economics on relatively Roman laws. We are, they argue, really much the same at base, governed by the same motivations and the same laws of human nature. Gary Becker is a notorious limiting-case. Becker won the Nobel Prize for modeling great swathes of what we do in day-to-day life under the principles of market equilibrium and rational choice theory, from drug addiction to racial discrimination to crime and family relations. Basically Becker supposes that the agents he models act so as to maximize their expected utility. The trick is to prescribe just what in the case under study utility consists in, which can include anything from financial gains to inconvenience to serious illness or the joys of watching your spouse consume. As you will see, I shall call this "climbing down the ladder of abstraction". Note that in Becker's cases this enterprise is relatively unconstrained, so the accounts are unfalsifiable, which many of us still take to be a damming charge. As economist Robert Pollak argues, "The devil is in the details." (2003, 120)

> Pischke and Angrist seem to have an optimistic view about breadth:
>
> . . . anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future. (2010, 23)

As I remarked, I am suspicious about principles of behavior that are supposed to apply almost across the board. But that is not the source of my worries about ladders. After all, even though the specific causal principles describing the functioning of the Cuisinart, the Dualit, and the Krups toasters are all different, still I agree that there are a set of even more basic principles that all three share. Even assuming shared principles and laying aside worries about falsifiability, trouble looms: There may be a set of laws that enable X to be a contributing cause to Y in the study and these laws may be shared with the target but in the target they do not connect X and Y. That's because what counts as a realization of a given factor in the study often cannot do so in the target.

This problem arises because of the way properties at different levels of abstraction piggyback on one another. To use vocabulary familiar from another problem area, abstract features are generally multiply realizable at the concrete level, but the abstract does not supervene on the concrete. The causes in a causal principle can be more or less abstract; because of the piggybacking, principles involving factors at different levels can all obtain at once. On a sphere, "The trajectories of bodies moving subject only to inertia are great circles" is true; so too is "The trajectories of bodies moving subject only to inertia are geodesics (ie, the shortest distance between two points)". They are equally true

because on a sphere, being a great circle *is* to be a geodesic.[13] For spheres there's a "ladder" down from the abstract "geodesic" to the more concrete "great circle", but there is no such ladder for Euclidean surfaces.

Generally the higher the level of abstraction of a causal principle, the more widely it is shared across populations. Bodies on Euclidean planes subject only to inertia follow geodesics but not great circles. And the lower the level, the more likely that the principle is only locally true. This can make serious problems when it comes to the stretch of the principles that RCTs can establish. The Bangladesh nutrition program provides a vivid example.

There was good evidence that the integrated nutrition program had worked in 20,000 Indian villages. But it failed on average in Bangladesh sites. Looking at the standard account of what went wrong, we will see that issues about levels of abstraction were at the heart. Nothing in this account supposes that Bangladeshis and Indians are altogether different. On the contrary it seems likely they share a common principle that allowed the program to improve children's nutrition in India. But this principle couldn't do the same job in Bangladesh because things in Bangladesh just aren't what they are in India.

I imagine those who adopted the program in Bangladesh expected Bangladesh and India to share a simple, common-sense principle:

> *Principle 1*: Better nutritional knowledge in mothers plus food supplied by the project for supplemental feeding improves the nutritional status of their children.

But they did not.

The first reason for the lack of impact in Bangladesh, it seems, was "leakage": The food supplied by the project was often not used as a supplement but as a substitute, with the usual food allocation for that child passing to another member of the family (STC 2003). The principle "Better nutritional knowledge in mothers plus food supplied by the project for supplemental feeding improves children's nutrition" was true in the original successful cases but not in Bangladesh. This suggest that a better shot at a shared principle would be:

> *Principle 2*: Better nutritional knowledge in mothers plus *supplemental feeding of children* improves children's nutrition.

This is a principle about features at a *higher level of abstraction* than those in the first principle. In the successful cases in India the more concrete feature "food supplied by the project" constituted the more abstract feature "supplemental feeding". But not in Bangladesh. There the ladders are missing that connect the abstract features in the shared principles with the concrete features offered by the program.

[13] I shall here be relatively cavalier about the metaphysics of properties. I treat abstract features and concrete ones both as real and I treat them as different features even if having one of these (the more concrete feature) is what constitutes having the more abstract one on any occasion. I take it that claims like this can be rendered appropriately, though probably differently, in various different metaphysical accounts of properties.

A second major reason for the lack of positive impact is also a problem with connecting ladders between the abstract and the concrete. It's labeled "the mother-in-law factor" by Howard White, who also points out what I call "the man factor":

> The program targeted the mothers of young children. But mothers are frequently not the decision makers . . . with respect to the health and nutrition of their children. For a start, women do not go to market in rural Bangladesh; it is men who do the shopping. And for women in joint households – meaning they live with their mother-in-law – as a sizeable minority do, then the mother-in-law heads the women's domain. Indeed, project participation rates are significantly lower for women living with their mother-in-law in more conservative parts of the country. (White 2009, 6)

This suggests yet another proposal for a shared principle:

> *Principle 3*: Better nutritional knowledge results in better nutrition for a child in those who
> a. provide the child with supplemental feeding
> b. control what food is procured
> c. control how food gets dispensed and
> d. hold the child's interests as central in performing b. and c.

Just as the food supplied by the project did not count as supplemental feeding in the Bangladesh program, mothers in that program did not in general satisfy the more abstract descriptions in b. and c.
The all-too-common fact that things in one setting may not be what they are in another makes real trouble for the use of RCTs as evidence. The previous successes of the program in India are relevant to predictions about the Bangladesh program only *relative to* the vertical identification of mothers with the more abstract features in b., c., and d. But not all of these identifications hold. So the previous successes are not evidentially relevant.

7. Roman laws, ladders, and structural parameters

The lesson of BINP is that the way abstract and concrete features relate implies:

> 1. In different contexts the same isn't always the same.

And,

> 2. This limits the usefulness of it-works-somewhere claims for predicting "It will work for us."

But the very same facts about the relations between the abstract and the concrete equally imply:

> 1.' In different contexts very different things can be the same.

And because of this,

> 2.' It-works-somewhere claims can support policy predictions in contexts far away and very different from the study populations that warrant them.

Pishke and Angrist employ this in their commendation of RCTs. "Small ball sometimes wins big games", they tell us

(2010, 25). How so? Because sometimes from RCTS, they urge, you can learn "structural econometric parameters", where following David Hendry, "Structure . . . is defined as the set of basic features of the economy which are invariant to [various specific] changes in that economy", including "an extension of the sample" (Hendry and Mizon 2010, 1-2). How wide an extension? That depends on the theory. For the moment let us assume, wide enough at least to cover the policy target.

Suppose that in the study a *structural* law of form L allows X to cause Y. Then β from that law is a structural parameter. Because β is a structural parameter, $\beta \neq 0$ in the study population shows that it's unequal to 0 in extensions of the population. This line of reasoning is familiar. Because the gravitational constant G is a structural parameter, Galileo can measure it on balls rolling down inclined planes and Euler a century later can put the same G into formulae calculating the "true curve" of cannonballs that are subject to the buoyant and resistant forces of the air as well as to gravity.

The parameter discussed by Angrist and Pischke is the "intertemporal [labor supply] substitution elasticity" (2010, 4); that is, a parameter that represents how much transitory wage changes contribute to hours of work a worker supplies. This is a theoretical parameter in, for example, life cycle theory. Is it constant enough for Angrist and Pishke to play the Galileo-Euler game? Maybe, maybe not. As Angus Deaton remarks, "Structural parameters are in the eye of the beholder."[14] Or see Mervyn King, Governor of the Bank of England: "There are probably few genuinely "deep" (and therefore stable) parameters or relationships in economics . . . ."[15]

I don't know if the labor supply elasticity is a structural parameter nor how far the structure stretches if it is. But Pischke and Angrist must take it that way. Here is the longer passage from which I quoted before:

> Small ball sometimes wins big games. In our field, some of the best research designs used to estimate labor supply elasticities exploit natural and experimenter-induced variation in specific labor markets. Oettinger . . . analyzes stadium vendors' reaction to wage changes driven by changes in attendance, while Fehr and Goette . . . study bicycle messengers in Zurich who, in a controlled experiment, received higher commission rates for one month only. (2010, 25)

Oettinger's analysis of stadium vendors at major-league baseball games (Oettinger 1999) supposes that the vendors' expectations about the size of the crowd constitute their wage expectations and in turn their wage expectations constitute "laborers" wage expectations' in this case. Similarly the number of vendors constitutes the labor supply in this case. So Angrist and Pishke seem to assume that labor supply elasticity is a structural parameter and that the parameter connecting vendor's expectations of crowd size with number of vendors showing up at the stadium *is* the labor supply elasticity in this situation.

What warrants these two assumptions? We confront here the twin problems of Roman laws and warranted ladders. For the first, it is usually theory that teaches that there is a structural parameter, but it had best be credible well-supported theory. As to the second, we need help both in climbing up the ladder of abstraction in the study situation; then in new settings, in climbing down. How do we know that what Oettinger measured on his stadium vendors was an instantiation of the labor supply parameter? And when we turn to a new situation with this parameter in hand, how do we figure what concrete features count as labor supply elasticity there? Theory can help. But it will also take sound knowledge of the local context. The point is that studies like Galileo's and Oettinger's – and RCTs – can *measure* structural parameters but they cannot tell us that that there is a structural parameter to be measured. That information must come from elsewhere.

8. Unbroken Bridges

My final problem involves *causal chains*. Generally getting from cause to effect is not a one-step process. Rather the policy variable is at the head of a causal chain with the hoped for outcome at the tail, with a number of links in between. Policy X causes outcome Y in the study situation because X causes U which causes V which causes W which . . . which causes Y. We can expect X to cause Y in a different situation only so long as the chain is unbroken.

Consider figure 1 and look at the first step: What enables X to cause U? I have been

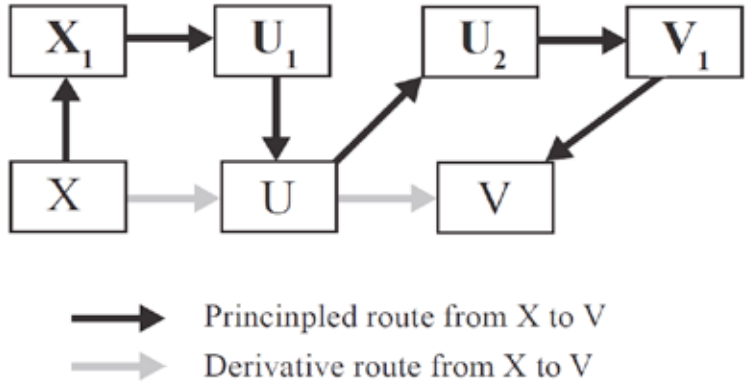


Figure 1: Two routes from a cause to an effect, at different levels of abstraction.

[14] Private conversation.

[15] Read at the Royal Society 22 March 2010.

arguing that it is often not because of a general principle connecting X and U but rather because X and U are concretizations of features X_1 and U_1 at a higher level of abstraction where X_1 and U_1 are joined by a reasonably general principle. Similarly U may cause V not because of a principle connecting U and V but rather because of a general principle between more abstract features U_2 and V_1 that they instantiate. Note the new subscripts. There is no reason that the very same features under which U is the effect of X should be the feature in virtue of which U is the cause of V.

Let me illustrate with a possible example social workers have been worrying about from UK child welfare policy where a child's care-givers are heavily encouraged, perhaps badgered, into attending parenting classes. It is illustrated in figure 2:

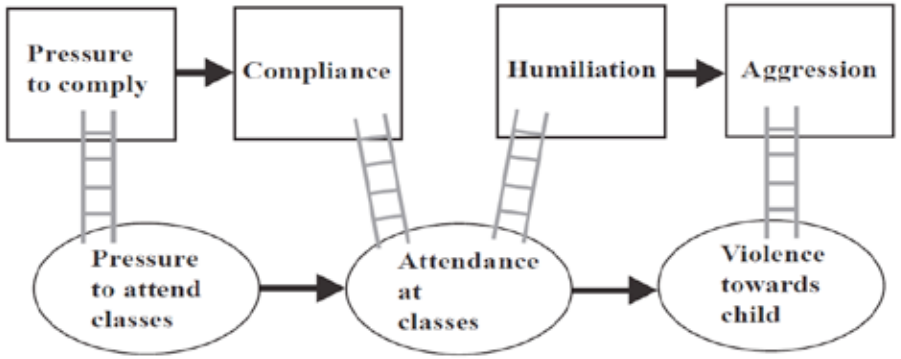


Figure 2: An example of a "broken bridge".

Consider making fathers attend parenting classes. Different cultures in the UK have widely different views about the roles fathers should play in parenting. Compelling fathers to attend parenting classes can instantiate the more abstract feature, "ensuring care-givers are better informed about ways to help the child" in which case it can be expected to be positively effective for improving a child's welfare. But it may also instantiate the more abstract feature "public humiliation", in which case it could act oppositely. Attending classes, as a result of pressure, can constitute a public humiliation and by virtue of being a public humiliation can lead to aggressive and violent behavior, which may be directed towards the child. There is then no unbroken bridge at the level of more widely applicable principle but there is a linked-up sequence at the more concrete level.

This of course has mixed policy implications. If we found that pressing fathers to attend parenting classes in this cultural group led to negative outcomes, that would not mean it should be expected to do so in other groups. The general principles that affect the different populations may be the same but they don't make an unbroken bridge for the negative effects to move along. On the other hand, getting positive results in other groups where the humiliation mechanism is not activated does not tell us what will be the overall outcome where it is activated. This is yet another case where knowing that a policy works – or fails – somewhere is at best a starting point for figuring out if it will work for us.

9. Conclusion

We can do better at predicting policy effectiveness. And philosophy helps show how. RCTs can help too, as their advocates maintain. But, as I have argued, it is a long and tortuous road from learning that a policy works somewhere, which is the kind of claim an RCT can clinch, to correctly predicting that it will – or won't – work for you. And you can go wrong in both directions: accepting programs that won't work for you, as Levy claims has repeatedly happened with Progresa, and rejecting ones that would, like the J-PAL rejection of textbooks in favor of deworming or in my hypothesized example, sending caregivers who won't feel humiliated to parenting classes.

I've rehearsed four essential materials it takes to secure a safe pathway:

1. Shared laws.
2. Supports.
3. Ladders.
4. Laws that interlock.

No matter how secure the starting point, if any one of these is missing, you just can't get there from here.

I don't need to remind you that a conclusion is only as secure as its weakest premise. RCTs may be gold standard for underpinning the start point but you can't pave the road in between with gold bricks. Evidence for these other factors is necessarily different and varied in form: theory – big and little, consilience of inductions, and a great deal of local information about study and target situations. Philosophy matters because once you know what you need, you can hunt for

it. And often you can find it. Here is Howard White again: "In the Bangladesh case, identification of the "mother-in-law" effect came from reading anthropological literature . . . ." (2009, 15) But to find it you must be encouraged to look. And where it doesn't exist, the sciences must be encouraged to uncover it. It's no good just putting all your money into gold bricks.

We philosophers of science are faced then with a hard job. Here as elsewhere in the natural and social sciences, in policy, and in technology, we can help. But to do so we need to figure out how better to engage with scientific practice and not just with each other.

References

Angrist, Joshua, and Jörn-Steffen Pischke. 2010. 'The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics.' *Journal of Economic Perspectives* 24: 3-30.

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

——— 2009. 'Causal Laws, Policy Predictions and the Need for Genuine Powers.' In *Dispositions and Causes*, ed. Toby Handfield, 127-158. Oxford: Oxford University Press.

Deaton, Angus. 2009. 'Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development.' *Proceedings of the British Academy 162: 123-160.*

Deaton, Angus. 2010. 'Instruments, Randomization, and Learning about Development.' *Journal of Economic Literature* 48: 424-455.

Duflo, Esther, and Michael Kremer. 2005. 'Use of Randomization in the Evaluation of Development Effectiveness.' In *Evaluating Development Effectiveness*, ed. George Pitman, Osvaldo Feinstein, and Gregory Ingram, 205-232. New Brunswick, NJ: Transaction Publishers.

Hendry, David and Grayham Mizon. 2010. 'Econometric Modelling of Changing Time Series.' Oxford University, Discussion Paper Series.

——— 2011. 'What Needs Rethinking in Macroeconomics?' *Global Policy* 2: 176-183.

The Lancet. 2004. 'The World Bank is finally embracing science.' *The Lancet* 364: 731-732.

Leamer, Edward. 2010. 'Tantalus on the Road to Asymptotia.' *Journal of Economic Perspectives* 24: 31-46.

Lucas, Robert. 1976. 'Econometric Policy Evaluation: A Critique.' in *The Phillips Curve and Labor Markets*, ed. Karl Brunner and Allan Meltzer. Amsterdam: North Holland.

Ludwig, Jens, et al. 2008. 'What Can We Learn about Neighborhood Effects from the Moving To Opportunity Experiment?' *American Journal of Sociology* 114: 144-188.

Mackie, John Leslie. 1965. 'Causes and Conditions.' *American Philosophical Quarterly* 2: 245-64.

Mill, John Stuart. 1836/1967. 'On the definition of political economy and on the method of philosophical investigation in that science.' In *Collected Works of John Stuart Mill*, vol. 4. Toronto: University of Toronto Press.

——— 1843/1850. *A System of Logic*. New York: Harper and Brothers.

Oettinger, Gerald. 1999. 'An Empirical Analysis of the Daily Labor Supply of Stadium Vendors." *Journal of Political Economy* 107: 360-392.

Pollak, Robert. 2003. 'Gary Becker's Contributions to Family and Household Economics.' *Review of Economics of the Household*, 1: 111-141.

SIGN. 2011. *SIGN 50: A guideline developer's handbook*. Edinburgh: Scottish Intercollegiate Guidelines Network.

STC. 2003. *Thin in the Ground: Questioning the evidence behind World Bank-funded community nutrition projects in Bangladesh, Ethiopia and Uganda*. London: Save the Children UK.

USDE. 2003. *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington, DC: Coalition for Evidence-Based Policy.

White, Howard. 2009. 'Theory-Based Impact Evaluation: Principles and Practice.' *3ie Working Paper* 3. New Delhi: International Initiative for Impact Evaluation.

World Bank. 1995. *Tamil Nadu and Child Nutrition: A New Assessment*. Washington, DC: World Bank.